21 al 25 de octubre de 2013
Mar del Plata

CACIC2013

XIX Congreso Argentino de Ciencias de la Computación

LIBRO DE ACTAS

XIX Congreso Argentino de Ciencias de la Computación - CACIC 2013 : Octubre 2013, Mar del
Plata, Argentina : organizadores : Red de Universidades con Carreras en Informática RedUNCI, Universidad CAECE / Armando De Giusti ... [et.al.] ; compilado por Jorge Finochietto ; ilustrado por María Florencia Scolari. - 1a ed. - Mar del Plata : Fundación de Altos Estudios en Ciencias Exactas, 2013.
E-Book.

ISBN 978-987-23963-1-2

1. Ciencias de la Computación. I. De Giusti, Armando II. Finochietto, Jorge, comp. III. Scolari, María Florencia, ilus.
CDD 005.3

Fecha de catalogación: 03/10/2013

AUTORIDADES DE LA REDUNCI

Coordinador Titular
De Giusti Armando (UNLP) 2012-2014

Coordinador Alterno
Simari Guillermo (UNS) 2012-2014

Junta Directiva
Feierherd Guillermo (UNTF) 2012-2014
Padovani Hugo (UM) 2012-2014
Estayno Marcelo (UNLZ) 2012-2014
Esquivel Susana (UNSL) 2012-2014
Alfonso Hugo (UNLaPampa) 2012-2013
Acosta Nelson (UNCPBA) 2012-2013
Finochietto, Jorge (UCAECE) 2012-2013
Kuna Horacio (UNMisiones) 2012-2013

Secretarias
Secretaría Administrativa: Ardenghi Jorge (UNS)
Secretaría Académica: Spositto Osvaldo (UNLaMatanza)
Secretaría de Congresos, Publicaciones y Difusión: Pesado Patricia (UNLP)
Secretaría de Asuntos Reglamentarios: Bursztyn Andrés (UTN)

CACIC2013

AUTORIDADES DE LA UNIVERSIDAD CAECE

Rector
Dr. Edgardo Bosch

Vicerrector Académico
Dr. Carlos A. Lac Prugent

Vicerrector de Gestión y Desarrollo Educativo
Dr. Leonardo Gargiulo

Vicerrector de Gestión Administrativa
Mg. Fernando del Campo

Vicerrectora de la Subsede Mar del Plata:
Mg. Lic. María Alejandra Cormons

Secretaria Académica:
Lic. Mariana A. Ortega

Secretario Académico de la Subsede Mar del Plata
Esp. Lic. Jorge Finochietto

Director de Gestión Institucional de la Subsede Mar del Plata
Esp. Lic. Gustavo Bacigalupo

Coordinador de Carreras de Lic. e Ing. en Sistemas
Esp. Lic. Jorge Finochietto

RedUNCI

UNIVERSIDAD CAECE 01 010 1961
UNIVERSIDAD CAECE
Mar del Plata

COMITÉ ORGANIZADOR LOCAL

Presidente
Esp. Lic. Jorge Finochietto

Miembros
Esp. Lic. Gustavo Bacigalupo
Mg. Lic. Lucia Malbernat
Lic. Analía Varela
Lic. Florencia Scolari
C.C. María Isabel Meijome
CP Mayra Fullana
Lic. Cecilia Pellerini
Lic. Juan Pablo Vives
Lic. Luciano Wehrli

Escuela Internacional de Informática (EII)
Directora
Dra. Alicia Mon

Coordinación
CC. María Isabel Meijome

COMITÉ ACADÉMICO

Universidad	Representante
Universidad de Buenos Aires	Echeverria, Adriana (Ingeniería) – Fernández
Universidad Nacional de La Plata	Slezak, Diego (Cs. Exactas)
Universidad Nacional del Sur	De Giusti, Armando
Universidad Nacional de San Luis	Simari, Guillermo
Universidad Nacional del Centro de la	Esquivel, Susana
Provincia de Buenos Aires	Acosta, Nelson
Universidad Nacional del Comahue	Vaucheret, Claudio
Universidad Nacional de La Matanza	Spositto, Osvaldo
Universidad Nacional de La Pampa	Alfonso, Hugo
Universidad Nacional Lomas de Zamora	Estayno, Marcelo
Universidad Nacional de Tierra del Fuego	Feierherd, Guillermo
Universidad Nacional de Salta	Gil, Gustavo
Universidad Nacional Patagonia Austral	Márquez, María Eugenia
Universidad Tecnológica Nacional	Leone, Horacio
Universidad Nacional de San Juan	Otazú, Alejandra
Universidad Autónoma de Entre Ríos	Aranguren, Silvia
Universidad Nacional Patagonia San Juan Bosco	Buckle, Carlos
Universidad Nacional de Entre Ríos	Tugnarelli, Mónica
Universidad Nacional del Nordeste	Dapozo, Gladys
Universidad Nacional de Rosario	Kantor, Raúl
Universidad Nacional de Misiones	Kuna, Horacio
Universidad Nacional del Noroeste de la Provincia de Buenos Aires	Russo, Claudia
Universidad Nacional de Chilecito	Carmona, Fernanda
Universidad Nacional de Lanús	García Martínez, Ramón

COMITÉ ACADÉMICO

Universidad	Representante
Universidad Nacional de Santiago del Estero	Durán, Elena
Escuela Superior del Ejército	Castro Lechstaler Antonio
Universidad Nacional del Litoral	Loyarte, Horacio
Universidad Nacional de Rio Cuarto	Arroyo, Marcelo
Universidad Nacional de Córdoba	Brandán Briones, Laura
Universidad Nacional de Jujuy	Paganini, José
Universidad Nacional de Río Negro	Vivas, Luis
Universidad Nacional de Villa María	Prato, Laura
Universidad Nacional de Luján	Scucimarri, Jorge
Universidad Nacional de Catamarca	Barrera, María Alejandra
Universidad Nacional de La Rioja	Nadal, Claudio
Universidad Nacional de Tres de Febrero	Cataldi, Zulma
Universidad Nacional de Tucumán	Luccioni, Griselda
Universidad Nacional Arturo Jauretche	Morales, Martín
Universidad Nacional del Chaco Austral	Zachman, Patricia
Universidad de Morón	Padovani, Hugo René
Universidad Abierta Interamericana	De Vincenzi, Marcelo
Universidad de Belgrano	Guerci, Alberto
Universidad Kennedy	Foti, Antonio
Universidad Adventista del Plata	Bournissen, Juan
Universidad CAECE	Finochietto, Jorge
Universidad de Palermo	Ditada, Esteban
Universidad Católica Argentina - Rosario	Grieco, Sebastián
Universidad del Salvador	Zanitti, Marcelo
Universidad del Aconcagua	Gimenez, Rosana
Universidad Gastón Dachary	Belloni, Edgardo
Universidad del CEMA	Guglianone, Ariadna
Universidad Austral	Robiolo, Gabriela

CACIC2013

COMITÉ CIENTÍFICO

Coordinación

Armando De Giusti (UNLP) - Guillermo Simari (UNS)

Abásolo, María José (Argentina)
Acosta, Nelson (Argentina)
Aguirre Jorge Ramió (España)
Alfonso, Hugo (Argentina)
Ardenghi, Jorge (Argentina)
Baldasarri Sandra (España)
Balladini, Javier (Argentina)
Bertone, Rodolfo (Argentina)
Bría, Oscar (Argentina)
Brisaboa, Nieves (España)
Bursztyn, Andrés (Argentina)
Cañas, Alberto (EE.UU)
Casali, Ana (Argentina)
Castro Lechtaller, Antonio (Argentina)
Castro, Silvia (Argentina)
Cechich, Alejandra (Argentina)
Coello Coello, Carlos (México)
Constantini, Roberto (Argentina)
Dapozo, Gladys (Argentina)
De Vicenzi, Marcelo (Argentina)
Deco, Claudia (Argentina)
Depetris, Beatriz (Argentina)
Diaz, Javier (Argentina)
Dix, Juerguen (Alemania)
Doallo, Ramón (España)
Docampo, Domingo
Echaiz, Javier (Argentina)
Esquivel, Susana (Argentina)
Estayno, Marcelo (Argentina)
Estevez, Elsa (Naciones Unidas)
Falappa, Marcelo (Argentina)
Feierherd, Guillermo (Argentina)
Ferreti, Edgardo (Argentina)
Fillottrani, Pablo (Argentina)
Fleischman, William (EEUU)
García Garino, Carlos (Argentina)
García Villalba, Javier (España)
Género, Marcela (España)
Giacomantone, Javier (Argentina)
Gómez, Sergio (Argentina)
Guerrero, Roberto (Argentina)
Henning Gabriela (Argentina)
Janowski, Tomasz (Naciones Unidas)
Kantor, Raul (Argentina)
Kuna, Horacio (Argentina)
Lanzarini, Laura (Argentina)
Leguizamón, Guillermo (Argentina)
Loui, Ronald Prescott (EEUU)
Luque, Emilio (España)
Madoz, Cristina (Argentina)
Malbran, Maria (Argentina)
Malverti, Alejandra (Argentina)
Manresa-Yee, Cristina (España)
Marín, Mauricio (Chile)
Motz, Regina (Uruguay)
Naiouf, Marcelo (Argentina)
Navarro Martín, Antonio (España)
Olivas Varela, José Ángel (España)
Orozco Javier (Argentina)
Padovani, Hugo (Argentina)
Pardo, Álvaro (Uruguay)
Pesado, Patricia (Argentina)
Piattini, Mario (España)
Piccoli, María Fabiana (Argentina)
Printista, Marcela (Argentina)
Ramón, Hugo (Argentina)
Reyes, Nora (Argentina)
Riesco, Daniel (Argentina)
Rodríguez, Ricardo (Argentina)
Roig Vila, Rosabel (España)
Rossi, Gustavo (Argentina)
Rosso, Paolo (España)
Rueda, Sonia (Argentina)
Sanz, Cecilia (Argentina)
Sposito, Osvaldo (Argentina)
Steinmetz, Ralf (Alemania)
Suppi, Remo (España)
Tarouco, Liane (Brasil)
Tirado, Francisco (España)
Vendrell, Eduardo (España)
Vénere, Marcelo (Argentina)
Villagarcia Wanza, Horacio (Arg.)
Zamarro, José Miguel (España)

RedUNCI

UNIVERSIDAD CAECE
Mar del Plata

CACIC2013

ACTAS DEL XIX CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN CACIC 2013

(Octubre 2013, Mar del Plata, Argentina)

XIV WORKSHOP AGENTES Y SISTEMAS INTELIGENTES - WASI -

XIII WORKSHOP PROCESAMIENTO DISTRIBUIDO Y PARALELO - WPDP -

XI WORKSHOP COMPUTACIÓN GRÁFICA, IMÁGENES Y VISUALIZACIÓN - WCGIV -

XI WORKSHOP TECNOLOGÍA INFORMÁTICA APLICADA EN EDUCACIÓN - WTIAE -

X WORKSHOP INGENIERÍA DE SOFTWARE - WIS -

X WORKSHOP BASES DE DATOS Y MINERÍA DE DATOS - WBDDM -

VIII WORKSHOP ARQUITECTURA, REDES Y SISTEMAS OPERATIVOS - WARSO -

V WORKSHOP INNOVACIÓN EN SISTEMAS DE SOFTWARE - WISS -

IV WORKSHOP ASPECTOS TEÓRICOS DE CIENCIA DE LA COMPUTACIÓN - WATCC -

IV WORKSHOP PROCESAMIENTO DE SEÑALES Y SISTEMAS DE TIEMPO REAL - WPSTR -

II WORKSHOP DE SEGURIDAD INFORMÁTICA - WSI -

II WORKSHOP DE INNOVACIÓN EN EDUCACIÓN EN INFORMÁTICA - WIEI -

III ETHICOMP LATINOAMÉRICA - ETH -

RedUNCI

UNIVERSIDAD CAECE 1967
UNIVERSIDAD CAECE
Mar del Plata

XIV WORKSHOP AGENTES Y SISTEMAS INTELIGENTES

- WASI -

CACIC2013

XIV WORKSHOP AGENTES Y SISTEMAS INTELIGENTES - WASI -

5653	**A Framework for Arguing from Analogy: Preliminary Results**	Maximiliano Budán (UNS), Paola Budán (UNSE), Guillermo Ricardo Simari (UNS)
5738	**An Approach to Argumentative Reasoning Servers with Conditions based Preference Criteria**	Juan Carlos Teze (UNS), Sebastián Gottifredi (UNS), Alejandro J. García (UNS), Guillermo Ricardo Simari (UNS)
5730	**On semantics in dynamic argumentation frameworks**	María Laura Cobo (UNS), Diego César Martínez (UNS), Guillermo Ricardo Simari (UNS)
5798	**Multi-criteria Argumentation-Based Decision Making within a BDI Agent**	Cecilia Sosa Toranzo (UNSL), Marcelo Errecalde (UNSL), Edgardo Ferretti (UNSL)
5665	**Una Extensión de Agentes en JASON para Razonar con Incertidumbre: G-JASON**	Adrian Biga (UNR), Ana Casali (UNR)
5850	**ABN: Considerando Características de los Objetos de Negociación**	Pablo Pilotti (CIFASIS), Ana Casali (UNR), Carlos Iván Chesñevar (UNS)
5657	**Dinámica de Conocimiento: Contracción Múltiple en Lenguajes Horn**	Nestor Jorge Valdez (UNCa), Marcelo A. Falappa (UNS)
5691	**Intelligent Algorithms for Reducing Query Propagation in Thematic P2P Search**	Ana Lucía Nicolini (UNS), Carlos Martín Lorenzetti (UNS), Ana Maguitman (UNS), Carlos Iván Chesñevar (UNS)

XIV WORKSHOP AGENTES Y SISTEMAS INTELIGENTES - WASI -

5862	**Red Pulsante con Aprendizaje Hebbiano para Clasificación de Patrones Ralos**	Iván Peralta (UNER), José T. Molas (UNER), César E. Martínez (UNL), Hugo L. Rufiner (UNER)
5786	**A Cognitive Approach to Real-time Rescheduling using SOAR-RL**	Juan Cruz Barsce (UTN-FRVM), Jorge Palombarini (UTN-FRVM), Ernesto Martínez (UTN-FRSF)
5878	**A Variant of Simulated Annealing to Solve Unrestricted Identical Parallel Machine Scheduling Problems**	Claudia R. Gatica (UNSL), Susana Esquivel (UNSL), Guillermo Leguizamón (UNSL)
5698	**Algoritmo evolutivo para el problema de planificación en proyectos de desarrollo de software**	Germán Dupuy (UNLPAM), Natalia Silvana Stark (UNLPAM), Carolina Salto (UNLPAM)
5879	**Algoritmos Evolutivos multire combinativos híbridos aplicados al problema de ruteo de vehículos con capacidad limitada**	Viviana Beatriz Mercado (UNPA), Andrea Villagra (UNPA), Daniel Pandolfi (UNPA), Guillermo Leguizamón (UNSL)
5733	**Análisis del comportamiento de un AG para GPUs**	Carlos Bermúdez (UNLPAM), Carolina Salto (UNLPAM)
5780	**A novel Competitive Neural Classiffier for Gesture Recognition with Small Training Sets**	Facundo Quiroga (UNLP), Leonardo Corbalán (UNLP)

A Framework for Arguing from Analogy: Preliminary Results

Paola D. Budán [1,2] Maximiliano C. D. Budán [1,2,3,4] and Guillermo R. Simari[2,3]

1 - Universidad Nacional de Santiago del Estero (UNSE)
2 - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)
3 - Universidad Nacional del Sur
4 - Concejo Nacional de Investigacion Científicas y Técnica (CONICET)
pbudan@unse.edu.ar-{mcdb,grs}@cs.uns.edu.ar

Abstract. Human reasoning applies argumentation patterns to draw conclusions about a particular subject. These patterns represent the structure of the arguments in the form of argumentation schemes which are useful in AI to emulate human reasoning. A type of argument schema is that what allow to analyze the similarities and differences between two arguments, to find a solution to a new problem from an already known one. Researchers in the heavily studied field of analogies in discourse have recognized that there is not a full and complete definition to indicate when two arguments are considered analogous. Our proposal presents an initial attempt to formalize argumentation schemes based on analogies, considering a relationship of analogy between arguments. This will contribute to the area increasing such schemes usefulness in Artificial Intelligence (AI), since it can be implemented later in Defeasible Logic Programming (DeLP).

1 Introduction

The ability to solve problems based on previous experience can be considered as an useful tool to develop and integrate to critical thinking. The act of thinking critically involves combining previous experiences with new experiences, finding patterns that follow those experiences, and considering the relationships among those patterns.

In the process of argumentation, information plays a fundamental role in supporting a point of view, making decisions, presenting the views of others, and solving new problems using past experiences. The human-like mechanism developed in computational argumentation research has made a significant contribution to the formalization of common sense reasoning and implementation of useful systems. In a general sense [15,7,2,14], argumentation can be associated with the interactive process where arguments for and against conclusions are offered, with the purpose of determining which conclusions are acceptable. Several argument-based formalisms have emerged, some of them based on Dung's seminal work called Abstract Argumentation Frameworks (AF)[5], others using

non-abstract or concrete forms of building arguments [1,7,6,12] leading to the application of these systems in many areas such as legal reasoning, recommender systems and multi-agent systems.

For Walton [18,17], Argumentation Schemes offer the possibility of representing the reasoning mechanisms on a semi-formal way, thus helping in the task of characterizing the inferential structures of arguments used in everyday discourse, particularly in special contexts such as scientific argumentation and AI systems in general. These simple devices capture the patterns of thought and expression from natural language, and contain questions that govern each of these patterns.

A particular type of argumentation scheme corresponds to *Argument from Analogy*, which represents a very common form of everyday human reasoning. In these schemes, two cases are analyzed for similarities and differences between them, using a form of inductive inference from a particular to a particular where the similarities between the cases lead to postulate a further similarity not yet confirmed; for instance, *"I have recently read H.G. Well's 'The Time Machine' and I liked it. Therefore, I will also like 'The War of the Worlds' by the same author"*. It should also evaluate if the perceived differences do not undermine the similarities between them. The argumentation from analogy allows to solve a new case based on already solved cases, or put it in a different way, to use previous experiences to consider a new case.

In this work, we will propose an extension of the abstract argumentation frameworks which allows to represent analogy between arguments determining the similarity degree or difference degree between them. This extension will be called *Analogy Argumentation Framework* (*AnAF*). This extension is motivated in the use of inferential mechanisms of argumentation based in the idea of argument from analogy that, as we said, are used in everyday situations in which a conclusion is obtained based on previous observations.

The paper is organized as follows. In section 2 is presented a brief introduction to argumentation schemes. Then, in Section 3 we introduce the concept of analogy. In Section 4, we present an introduction to argumentation framework. The core contribution of the paper is presented in Section 5 called as *Analogy Argumentation Framework*. Finally, in Section 6 we present the related work associated with the central issue of the work, and in Section 7 we conclude and propose future works.

2 Argumentation Schemes

There are several argumentation schemes proposed by Walton [18] applied to different areas such as in the legal and scientific communities, and in learning environments. These schemes are gaining importance in the field of AI, particularly because they allow the representation of defeasible arguments, *i.e.*, that can be refuted by who receives the argument, who thinks critically in relation to a given position. There are various argumentation schemes proposed by Walton [20], such as arguments coming from experts, from popular opinion, or from signs, among others.

In this paper, we focus on the *Argumentation from Analogy Scheme*. This scheme considers two cases C_1 and C_2 assessing the similarities and differences between them. The defeasible character is introduced by the specific differences between the cases C_1 and C_2. Walton defined three *critical questions* that are appropriate for using the scheme of argument from analogy:

1. Are there differences between C_1 and C_2 that would tend to undermine the force of the similarity cited?
2. Is the feature A true (false) in C_1?
3. Is there some other case C_3 that is also similar to C_1, but in which the feature A is false (true)?

In the words of Walton [18]: "*In general, the first critical question for the argument from analogy tends to be the most important one to focus on when evaluating arguments from analogy. If one case is similar to another in a certain respect, then that similarity gives a certain weight of plausibility to the argument from analogy. But if the two cases are dissimilar in some other respect, citing this difference tends to undermine the plausibility of the argument. So arguments from analogy can be stronger or weaker, in different cases.*"

In a recent work [19], Walton has analyzed different possibilities for this type of schema and has offered his understanding of how the schema integrates with the usage of argument from classification and the argument from precedent when applied in case-based reasoning by the use of a dialogue structure; below, we will summarily discuss these ideas. Next, we will focus on a detailed study of the concept of analogy, and define a relation of analogy between argument entities.

3 The Concept of Analogy

The term *analogy* has been widely studied as to their meaning and usage. Hesse [8], argues that the word is self-explanatory, and that two objects or situations are similar if they share some properties and differ in others. Walton [19] agrees with this perspective adding that two things are similar when they are visibly similar or they look similar. As to how to determine when two arguments are similar, Hesse uses a comparison between arguments based on the use of mathematical proportions. On the other hand, in a refinement of Hesse's idea, Walton points out that it is not easy to clearly define the comparison between arguments, as this requires interpreting the similarities and differences between them at various levels.

Offering another view, Carbonell [3] proposes a technique based on how we solve problems. This technique takes into account information from previous experience, which is useful for solving a new problem, as long as both occur in similar contexts; that is, the context of the problem determines a set of constraints under which the proposed solution is feasible. In [16], Sowa argues that it is possible to make a comparison between arguments, establishing a function of similarity or correspondence between them; and, by using another function,

referred to as the estimation function, it is possible to find the differences between the arguments. In a parallel effort, in [4] Cecchi *et al.* characterized and formalized relationships that capture the behaviour of a preference criterion among arguments; while this does not refer specifically to arguments from analogy, shows the usefulness in approaching the analogy between two arguments as a binary relationship.

These questions have received different answers and remains the focus of different research lines. Briefly, two objects or situations are analogous when they have some similar properties, maintaining other properties different. The similarity is then related to the properties shared between two objects or situations being compared. Following previous work, our proposal is to consider the analogy between two arguments A and B relaying on the following items defined next.

Definition 1 (Analogy Elements). *Given a set AR of arguments, we introduce:*

1. *a* constraint set*, denoted as Δ, contains the features governing the comparison of arguments in a given situation.*
2. *a* similarity degree *between two arguments A and B, denoted as $\alpha_\Delta(A, B)$, as a function: $\alpha_\Delta : AR \times AR \to [0, 1]$,*
3. *a* difference degree *between two arguments A and B, denoted as $\beta_\Delta(A, B)$, as a function: $\beta_\Delta : AR \times AR \to [0, 1]$,*

furthermore, for all $A, B \in AR$, it holds (1) $\alpha_\Delta(A, B) = \alpha_\Delta(B, A)$, (2) $\beta_\Delta(A, B) = \beta_\Delta(B, A)$, *and* (3) $\alpha_\Delta(A, B) + \beta_\Delta(A, B) = 1$.

The set Δ specifies the features that are significant to consider to establish whether two arguments are analogous or not. The content of this set is heavily dependent on the domain where the arguments are considered; thus, this is a semantic concept from which we will abstract away introducing the tools that will handle these features building the infrastructure for arguing from analogy. In the same way as Δ, the two functions α_Δ and β_Δ are dependent on the domain of application; therefore, although they remain unspecified in the formalization, a concrete definition must be given when implementing the framework. It is important to remark that in this initial approach, as the definition establishes, there is no difference in comparing A with B or B with A. This decision of not assigning preference to the features is a simplifying one, taken in the spirit of analyzing the simplest problem. In the future evolution of these ideas, we will to consider some form of preference over Δ's elements, and this preference will help in the comparison in a natural way introducing different possibilities.

Naturally, if between the arguments being compared the similarity degree is greater than the difference degree under the constraint set, it can be considered that the arguments are analogous; otherwise, differences prevail and they are considered as not analogous. Observe that the similarity degree and the difference degree are mutually dependent, *e.g.*, if the similarity degree between two arguments is 0.7, then the difference degree between them is 0.3. The following definition formalizes the analogy relation between arguments.

Definition 2 (Analogy Relation). *Let AR be a set of arguments and Δ be a constraint set. An* analogy relation*, denoted Γ_Δ, is defined as a binary relation on AR under the constraint set Δ, where the relation Γ_Δ is such that $\Gamma_\Delta \subseteq AR \times AR$ and satisfies the constrains of Δ, where $(A, B) \in \Gamma_\Delta$ iff $\alpha_\Delta(A, B) > \beta_\Delta(A, B)$, i.e., the similarity degree between them is greater than their difference degree.*

From the previous definition of analogy relation Γ_Δ, we can establish the following properties hold for analogy when arguments are compared over the same features or properties:

- *Reflexive*: $A \ \Gamma_\Delta \ A$. Any argument is analogous to itself.
- *Symmetric*: If $A \ \Gamma_\Delta \ B$, then $B \ \Gamma_\Delta \ A$. The analogy relation is symmetric by definition, *i.e.*, the analogy between arguments is established from the similarity between both, under a constraint set. This explains why the analogy relation is symmetric and is not asymmetrical.
- *Transitive*: If $A \ \Gamma_\Delta \ B$ and $B \ \Gamma_\Delta \ C$, then $A \ \Gamma_\Delta \ C$. The arguments A and B are similar according to the constraint set defined, the same occurs between the arguments B and C, thus A and C are similar and relation is transitive.

That is, the analogy relation between arguments under a given constraint set is an equivalence relation under that constraint set, and each equivalence class will contain all the arguments that have identical features in the frame of Δ. Also notice that the analogy relation is not equality since two argument that are analogous under a constraint set might no be analogous under a different constraint set. The definition of the analogy relation between arguments under a constraint set just introduced, will allow us to reformulate the questions for guiding the argumentation from analogy scheme, in the following way:

1. Are A and B analogous? Is $(A, B) \in \Gamma_\Delta$?
2. Are there differences between A and B that would tend to undermine the force of the similarity cited? Is $\alpha_\Delta(A, B) > \beta_\Delta(A, B)$?

4 Abstract Argumentation

Dung [5] introduced Abstract Argumentation Frameworks (AF) as an abstraction of a defeasible argumentation system. In an AF, an argument is an abstract entity with unspecified internal structure, and its role in the framework is solely determined by the attack relation it keeps with other arguments; thus, an AF is defined by a set of arguments and the attack relation defined over it.

Definition 3 (Argumentation Framework [5]). *An argumentation framework (AF) is a pair described as $\langle AR, Attacks \rangle$, where AR is a set of arguments, and the binary relation $Attacks \subseteq AR \times AR$.*

When it happens that $(A, B) \in Attacks$, we say that A attacks B, or that B is attacked by A. Likewise, extending the relation of attack, we will say that the set S attacks C when there exists at least an argument $A \in S$, such that

$(A, C) \in Attacks$. Given an AF, intuitively $A \in AR$ is considered *acceptable* if A can be defended of all its attackers (arguments) with other arguments in AR; this is formalized in the following definitions [5].

Definition 4 (Acceptability). *Let $AF = \langle AR, Attacks \rangle$ be a framework.*

- *A set $S \subseteq AR$ is said* conflict-free *if there are no arguments $A, B \in S$ such that $(A, B) \in Attacks$.*
- *$A \in AR$ is* acceptable *with respect to $S \subseteq AR$ iff for each $B \in AR$, if B attacks A then there exists $C \in S$ such that $(C, B) \in Attacks$; in such case it is said that B is attacked by S.*
- *A conflict-free set S is* admissible *iff each argument in S is acceptable with respect to S.*
- *An admissible set $S \subseteq AR$ is a* complete extension *of AF iff S contains every argument acceptable with respect to S.*
- *A set $S \subseteq AR$ is a* grounded extension *of AF iff S is a complete extension that is minimal with respect to set inclusion.*

We will now extend the Dung's framework introducing the possibility of taking in consideration the similarities and differences between arguments.

5 Analogy Argumentation Framework

Recently, the field of application of argumentation has been expanding, with the interesting addition of the research on argumentation schemes; however, still there is need to further formalize the structure of these schemes. Here, we will make a first approximation to this formalization through extending AFs to *Analogy Argumentation Frameworks (AnAF)*, introducing the consideration of the analogy between arguments in the well-known argument from analogy scheme, by representing the notions of similarities and differences between arguments.

When considering analogy among the set of arguments is natural, and intuitively appealing, to requiere two things: (1) that arguments that are analogous do not attack each other, and (2) if an argument attacks another, then any argument analogous to the attacker should be an attacker to same argument. This can easily formalized by taking advantage of the analogy relation that happens to be an equivalence relation. Let $[A] = \{X \in AR \,|\, X \Gamma_\Delta A\}$ be the class of arguments equivalent to A and AR_{Γ_Δ} the quotient set of AR by Γ_Δ.

Definition 5 (Nonconflicting Class). *Given an $AF = \langle AR, Attacks \rangle$, and an analogy relation Γ_Δ defined over AR. Let $[A] \in AR_{\Gamma_\Delta}$, $[A]$ is said to be a* nonconflicting class *iff there is no pair of arguments $X, Y \in [A]$ such that $(X, Y) \in Attacks$. The AR is said to be Γ_Δ-conformant iff all classes in the quotient set AR_{Γ_Δ} are non-conflicting.*

Definition 6 (Class Attack Relation). *Given an $AF = \langle AR, Attacks \rangle$, and an analogy relation Γ_Δ defined over AR. Let $[A] = \{X \in AR \,|\, X \Gamma_\Delta A\}$ be the class of arguments equivalent to A and AR_{Γ_Δ} the quotient set of AR by Γ_Δ. We say that Attacks is a class attack relation over AR_{Γ_Δ} iff when $A, B \in AR$,*

and $(A, B) \in Attacks$ *it happens that every argument* $X \in [A]$ *attacks every argument* $Y \in [B]$.

Definition 7 (AnAF). *Given* $AF = \langle AR, Attacks \rangle$*, and an analogy relation* Γ_Δ *defined over* AR*. An Analogy Argumentation Framework (AnAF) is a 3-tuple* $\Theta = \langle AR, Attacks, \Gamma_\Delta \rangle$ *where* AR *is a set of arguments, and* $Attacks$ *is a* Γ_Δ*-conformant, class attack relation.*

These definitions follow the intuitions expressed in (1) and (2) above.

Given an $AnFA$, an argument A is considered *Analogy-acceptable* if it can be defended of all its attackers (arguments) with other arguments in AR.

Definition 8 (Analogy Acceptability). *Let* $\Theta = \langle AR, Attacks, \Gamma_\Delta \rangle$ *be an* $AnAF$*. The acceptability of* $S \subseteq AR$ *is given by the following conditions:*

- $S \subseteq AR$ *is an Analogy-Conflict-Free set iff there are no arguments* $A, B \in S$ *such that* $(A, B) \in Attacks$*.*
- $A \in AR$ *is an Analogy-Acceptable with respect to* $S \subseteq AR$ *iff for each* $B \in AR$*, if* B *attacks* A *then exist* $C \in S$ *such that* B *is attacked by* C*.*
- *An Analogy-Conflict-Free set* S *is Analogy-Admissible iff each argument in* S *is Analogy-Acceptable with respect to* S*.*
- *An Analogy-Admissible set* $S \subseteq AR$ *is an Analogy-complete extension of* Θ *iff* S *contains each argument that is Analogy-Acceptable with respect to* S*.*
- $S \subseteq AR$ *is the Analogy-grounded extension of* Θ *iff* S *is an Analogy-complete extension that is* $\subseteq$*-minimal.*

Example 1 *Consider a scenario where an agent must decide whether it is riskier to invest in a real estate property or to invest in gold bullion, to reach a decision the agent ponders these arguments:*
A: I should invest my savings in real estate because they do not depreciate quickly, and this leads financial safety.
B: It is better to invest in gold bullion because it does not deteriorate, and it does not require maintenance as real estate does. It is not wise to invest in real estate because they lose value in many ways.
C: Investing in gold bullion is expensive because you have to store them in a safekeeping place. Land does not deteriorate, does not depreciate fast, does not require a place to store and provide financial reinsurance. I should not invest in gold bullion.
D: Buying land is a good way to invest whenever you carefully look for a place. Land does not devalue easily.
E: Buying foreign currency is an investment of unpredictable results because it depends on the global economy.
Let $\Theta = \langle AR, Attacks, \Gamma_\Delta \rangle$ *be an AnAF, where:*$AR = \{A; B; C; D; E\}$, $Attacks = \{(B, A); (C, B); (B, D)\}$ $\Gamma_\Delta = \{(A, C); (D, A); (D, C); (C, A); (A, D); (C, D)\}$, *and* Δ *is "Invest the savings into something that is not quickly devalued", and the similarity and difference degrees are represented in table 1.*
Some of analogy-conflict-free set are: $S_1 = \{A\}$, $S_2 = \{A; C\}$, $S_3 = \{A; C; D\}$, $S_4 = \{B\}$ *and* $S_5 = \{B\}$*. Note, for example, the set* $S_6 = \{A; E\}$ *is a conflict-free, but is not an analogy-conflict-free given that there is not an analogy relation*

$\alpha_\Delta(B,A) = 0$ $\beta_\Delta(B,A) = 1$	$\alpha_\Delta(C,B) = 0$ $\beta_\Delta(C,B) = 1$	$\alpha_\Delta(A,E) = 0$ $\beta_\Delta(A,E) = 1$	$\alpha_\Delta(B,D) = 0$ $\beta_\Delta(B,D) = 1$	$\alpha_\Delta(C,E) = 0$ $\beta_\Delta(C,E) = 1$
$\alpha_\Delta(A,C) = 1$ $\beta_\Delta(A,C) = 0$	$\alpha_\Delta(D,A) = 1$ $\beta_\Delta(D,A) = 0$	$\alpha_\Delta(B,E) = 0$ $\beta_\Delta(B,E) = 1$	$\alpha_\Delta(D,C) = 1$ $\beta_\Delta(D,C) = 0$	$\alpha_\Delta(E,D) = 0$ $\beta_\Delta(E,D) = 1$

Fig. 1: Coefficients Table

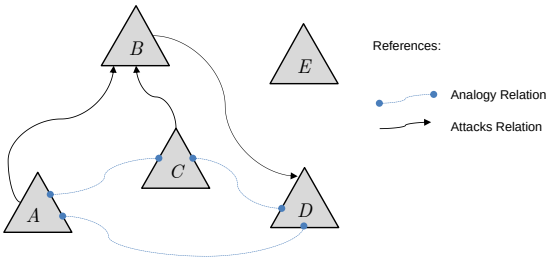


Fig. 2: Example for AnAF

between them on the constraint set considered.
The argument $A \in AR$ is Analogy-Acceptable with respect to S_3, since $B \in AR$ and $(B, A), (C, B) \in$ Attacks, and $C \in S_3$. Additionally, $(C, B) \notin \Gamma_\Delta$ and $(A, C) \in \Gamma_\Delta$. The argument $B \in AR$ is not Analogy-Acceptable because $(C, B) \in$ Attacks and there is no an argument that attacks C.
The set S_3 is an Analogy-Admissible because each argument in S_3 is Analogy-Acceptable in S_3. Additionally, S_3 is an Analogy-complete extension of Θ, and S_3 is an Analogy-grounded extension of Θ.

6 Related Works

Few studies exist formalizing the argumentation schemes proposed by Walton. However, there are several extensions of Dung's framework that are inspiring for this paper. Prakken [12] proposed Argumentation Systems with Structured Arguments, which used the structure of arguments and external preference information to define the a defeat relation. In this paper, we use the term "defeat" instead of the "attack", because defeat allows to considerer an attack relation plus preferences. Regarding argumentation schemes, Prakken [13] proposes that modeling reasoning using argumentation schemes necessarily involves developing a method combining issues of non-monotonic logic and dialogue systems. Nielsen *et al.* [11] claim that Dung's framework is not enough to represent argumentation systems with joint attacks, and they generalize it allowing a set of arguments to attack on a single argument. Modgil [10] also extends Dung's framework, preserving abstraction and expressing the preference between arguments. To do this, incorporates a second attack relation that characterizes the preference between arguments. Regarding to preference relation between arguments Cecchi*et al.* [4] defined this as a binary relation considering two particular criteria, specificity and equi-specificity, together with priorities between rules, defining preferred arguments and incomparable arguments.

In regards specifically to formalizing argumentation schemes, Hunter [9] presented a framework for meta-reasoning about object-level arguments allowing the presentation of richer criteria for determining whether an object-level argument is warranted. These criteria can use meta-information corresponding to the arguments, including the proponents and their provenances, and an axiomatization using this framework for reasoning about the appropriated conduct of the experts that introduce them. He shows how it can conform to some proposed properties for expert-based argumentation describing a formal approach to modelling argumentation providing ways to present arguments and counterarguments, and evaluating which arguments are, in a formal sense, warranted. He proposed a way to augment representation and reasoning with arguments at the object-level with a meta-level system for reasoning about the object-level arguments and their proponents. The meta-level system incorporates axioms for raising the object-level argumentation to the meta-level (an important case is to capture when an argument is a counterargument for another argument), and meta-level axioms that specify when proponents are appropriated for arguments. The meta-level system is an argumentation system to the extent that it supports the construction and comparison of meta-level arguments and counterarguments.

7 Conclusions and future works

Human reasoning applies argumentation patterns to draw conclusions about a particular subject. These patterns represent the structure of the arguments in the form of argumentation schemes which are useful in AI to emulate human reasoning. Argumentation schemes are a semiformal way of representing reasoning patterns. In this paper we presented an extension of Dung's frameworks, called *Analogy Argumentation Framework (AnAF)*, which allows to consider the similarity and difference degrees between two arguments in the context of an analogy relation. The analogy between arguments allows to approach the solution of a new case based on already solved cases, or put it in another way, to re-use previous experiences. The analogy relation represents in this proposal a form of a preference between arguments. As work in progress, we analyzed and studied the extensions of the classical semantics proposed by Dung within this new framework. It seems also necessary to formalize other argumentation schemes [20].

As future work, we will develop an implementation of the application of *AnAF* in the existing Defeasible Logic Programing system [1] as a basis. For doing that we will decrease the level of abstraction studying the internal structure of the arguments. To get the similarity degree between arguments involves implementing a mapping function between them, subject to a given constraint set, while determining the difference degree requires to implement a function to estimate differences between the arguments in question, subject to the same constraint set. The similarities or differences between arguments is in fact to compare premises, conclusions or inference mechanisms between arguments. The result-

[1] See http://lidia.cs.uns.edu.ar/delp

ing implementation will be exercised in different domains requiring to model analogy between arguments.

References

1. P. Besnard and A. Hunter. A Logic-Based Theory of Deductive Arguments. *Artif. Intell.*, 128(1-2):203–235, 2001.
2. P. Besnard and A. Hunter. *Elements of argumentation*, volume 47. MIT press, 2008.
3. J.G. Carbonell. *Learning by analogy: Formulating and generalizing plans from past experience*. Springer, 1983.
4. L. Cecchi and G.R. Simari. Sobre la relación de preferencias entre argumentos. In *VIII Congreso Argentino de Ciencias de la Computación*, 2002.
5. P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.
6. P.M. Dung, R.A. Kowalski, and F. Toni. Assumption-based argumentation. In I. Rahwan and G. R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 198–218. Springer, 2009.
7. A. J. García and G. R. Simari. Defeasible Logic Programming: An Argumentative Approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.
8. M.B. Hesse. *Models and analogies in science*, volume 7. University of Notre Dame Press Notre Dame, 1966.
9. A. Hunter. Reasoning about the appropriateness of proponents for arguments. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence, USA*, 2008.
10. S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9):901–934, 2009.
11. S. H. Nielsen and S. Parsons. A generalization of Dung's abstract framework for argumentation: Arguing with sets of attacking arguments. In *Argumentation in Multi-Agent Systems*, pages 54–73. Springer, 2007.
12. H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.
13. H. Prakken. On the nature of argument schemes. *Dialectics, dialogue and argumentation. An examination of douglas Waltons theories of reasoning and argument*, pages 167–185, 2010.
14. I. Rahwan and G.R. Simari. *Argumentation in artificial intelligence*. Springer, 2009.
15. G. R. Simari and R. P. Loui. A Mathematical Treatment of Defeasible Reasoning and its Implementation. *Artificial Intelligence*, 53(1–2):125–157, 1992.
16. J.F. Sowa and A.K. Majumdar. Analogical reasoning. In *Conceptual Structures for Knowledge Creation and Communication*, pages 16–36. Springer, 2003.
17. D. Walton. Justification of argumentation schemes. *Australasian journal of logic*, 3:1–13, 2005.
18. D. Walton. *Fundamentals of critical argumentation*. Cambridge Univ Press, 2006.
19. D. Walton. Similarity, precedent and argument from analogy. *Artificial Intelligence and Law*, 18(3):217–246, 2010.
20. D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, Cambridge, UK, 2008.

An Approach to Argumentative Reasoning Servers with Conditions based Preference Criteria

Juan Carlos Teze[1,2,3], Sebastián Gottifredi[1,3],
Alejandro J. García[1,3] and Guillermo R. Simari[1]

[1]Artificial Intelligence Research and Development Laboratory (LIDIA)
Department of Computer Science and Engineering (DCIC)
Universidad Nacional del Sur (UNS) - Alem 1253
(8000) Bahía Blanca, Buenos Aires, Argentina

[2]Agents and Intelligent Systems Area, Fac. of Management Sciences,
Universidad Nacional de Entre Ríos (UNER)
(3200) Concordia, Entre Ríos, Argentina

[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
e-mail:{jct,sg,ajg,grs}@cs.uns.edu.ar

Abstract. Argumentation is a reasoning mechanism attractive due to its dialectical and non monotonic nature, and its properties of computational tractability. In dynamic domains where the agents deal with incomplete and contradictory information, an argument comparison criterion can be used to determine the accepted information. Argumentation systems with a single argument comparison criterion have been widely investigated; in some of these approaches, the comparison criterion is fixed while in others a criterion can be selected and replaced in a modular way. We present an argumentative server providing client agents with recommendations and giving the possibility of specifying which of the available argument comparison criteria will be used to answer a query; for that, we formalize a special type of contextual query which by the use of conditions allows the server to dynamically change the criterion providing a declarative way of representing users preferences.

1 Introduction

A defeasible argumentation system provides ways of confronting contradictory statements to determine whether some particular information can be accepted or warranted [9,1,6,7]. To obtain an answer, an argumentation process goes through a series of steps. A very important one is the comparison of conflicting arguments to decide which one prevails. The definition of a formal comparison criterion thus becomes a central problem in defeasible argumentation.

Argumentation systems using a single argument comparison criterion have been widely studied in the literature [13,16,18,3,17,9]. The comparison criterion represents a fundamental part of an argumentation system because the inferences

an agent can obtain from its knowledge will depend on the criterion being used. In some of these approaches, the comparison criterion is fixed while in others a criterion can be selected and replaced in a modular way. The main contribution is to provide a more declarative way of representing users preferences by means of a framework where several comparison criteria can be taken into account and the selection of one depend on the specific conditions that are satisfied. Next, we present an example that will serve two purposes: to motivate the main ideas of our proposal, and to serve as a running example.

Example 1 *Consider an on_board_computer making recommendations to the user of the vehicle in which is installed, like suggesting a hotel. To give advice, the computer uses two types of knowledge: the user's particular preferences which are obtained before starting the travel; and particular information about the user's context, obtained dynamically during the travel. Besides this knowledge, the computer also needs certain criteria for making decisions to give recommendations. The computer could use some of the following criteria: security and comfort. However, the driver can restrict the criteria that the computer may use by means of conditions; for instance, before starting a journey the driver may tell the computer that if the road is blocked by striking workers, the computer has to consider the security criteria. In this situation, the restrictions are defined over the possible existence of specific knowledge stored in the computer and, depending on the chosen criteria, the user might receive contradictory recommendations.*

Recently, there have been important developments in AI regarding contextual and conditional preferences [5,2]. A particularly active area is focused on the association of conditions to users preferences [4,12]. In [4], a conditional preference network (*CP-nets*) is proposed. Like us, the authors present a model for representing and reasoning with the user preferences, where conditional preference expressions are permitted. In contrast to [4], our approach is defined over dynamic domains where the agents deal with incomplete and contradictory information. We formalize a special type of contextual query which by means of conditions allows the server to dynamically know what criterion to choose.

The study of Recommender Systems [14,15,11,8] has become important in AI over the last decade. We focus on a particular form of implementing recommender systems, called Recommender Servers that extends the integration of argumentation and recommender systems to a Multi-Agent System setting. Recommender Servers are based on an implementation of DeLP [9] called DeLP-Server [10]. In this paper we will introduce a defeasible logic programming recommender server which allows the clients to select through conditional expressions which criteria the server will use to answer their queries.

The rest of the paper is structured as follows. In Section 2 we will present the necessary background introducing basic definitions and some works that will be used in the rest of the paper; then, in Section 3 we will introduce the recommender server whose reasoning will be addressed by one of the comparison criteria indicated in the client query. To illustrate the formalism, in Section 4 we introduce an example in DeLP. Finally, in Section 5 we discuss related work and offer our conclusions and the possible directions for our future work.

2 Preliminary Background

In [10], an extended implementation of DeLP, called DeLP-server, has been presented; this system provides an argumentative reasoning service for multi-agent systems. A DeLP-server is a stand-alone application that stores a DeLP-program that is used to answer client queries. To answer queries, the DeLP-server will use the public knowledge stored in it as a Defeasible Logic Program, complementing this with individual knowledge a client agent might send, thus creating a particular scenario for the query (see Fig. 1). This information modifying the public knowledge stored in the DeLP-server is called context, and denoted $\mathcal{C}o$.

In DeLP, knowledge is represented using facts, strict and defeasible rules. *Facts* are ground literals representing atomic information, or the negation of atomic information using the strong negation "$\sim$". An overlined literal will denote the complement of that literal with respect to strong negation, *i.e.,* $\overline{L}$ is $\sim L$, and $\overline{\sim L}$ is L. *Strict Rules* are denoted $L_0 \leftarrow L_1, \ldots, L_n$ and represent firm information, whereas *Defeasible Rules* are denoted $L_0 \prec L_1, \ldots, L_n$ and represent defeasible knowledge, *i.e.,* tentative information, where the head L_0 is a literal and the $body\{L_i\}_{i>0}$ is a set of literals. A Defeasible Logic Program $\mathcal{P}$ (DeLP-program for short) is a set of facts, strict rules and defeasible rules.

Example 2 *Continuing with Ex. 1, let $\mathcal{P}_l$ be a DeLP-program that models the information stored inside the on_board_computer and $\mathcal{P}_c$ be a DeLP-program representing the private pieces of information related to a driver particular context:*

$$\mathcal{P}_l = \left\{ \begin{array}{l} nearby\_lodging \prec included\_AC. \\ \sim nearby\_lodging \prec dangerous\_area. \\ road\_blocked \leftarrow workers\_on\_strike. \\ dangerous\_area \leftarrow zone\_lots\_of\_thefts. \\ downtown \leftarrow zone\_lots\_of\_restaurants. \end{array} \right\} \quad \mathcal{P}_c = \left\{ \begin{array}{l} zone\_lots\_of\_thefts. \\ zone\_lots\_of\_restaurants. \\ included\_AC. \\ workers\_on\_strike. \end{array} \right\}$$

Given a DeLP-program $\mathcal{P}$, a derivation for a literal L from $\mathcal{P}$ is called 'defeasible', because there may exist information that contradicts L.

Definition 1 (Defeasible/Strict Derivation) *[9] Let $\mathcal{P}$ be a DeLP-program and L a ground literal. A defeasible derivation of L from $\mathcal{P}$, denoted $\mathcal{P} \mid\sim L$, is a finite sequence of ground literals $L_1, L_2, \ldots, L_n = L$, where each literal L_i is in the sequence because:*

(a) L_i is a fact in $\mathcal{P}$, or
(b) there exists a rule R_i in $\mathcal{P}$ (strict or defeasible) with head L_i and body $B_1, B_2, \ldots, B_k$ and every literal of the body is an element L_j of the sequence appearing before L_i $(j < i)$.

We will say that L has a strict derivation *from $\mathcal{P}$, denoted $\mathcal{P} \vdash L$, if either L is a fact or all the rules used for obtaining the sequence $L_1, L_2, \ldots, L_n$ are strict rules.*

Two literals are contradictory if they are complementary. Let $\mathcal{P}$ be a DeLP-program, the program $\mathcal{P}$ is coherent iff there are no strict derivations for two contradictory literals from $\mathcal{P}$. The set of literals supported by strict derivations is assumed to be non-contradictory, since these derivations are based on strict rules, which cannot be defeated.

A query is a literal Q, and the set of all possible queries will be denoted $\mathbb{Q}$. In [10], several contextual queries were defined, these types of queries allow the inclusion of private pieces of information related to the agents's particular context that will be taken into consideration when computing the answers.

Definition 2 (Contextual query) *[10]*
Given a DeLP-program $\mathcal{P}$, a contextual query for $\mathcal{P}$ is a pair [Ls, Q] where Ls is a non-contradictory set of literals, and Q is a query.

Three operators for DeLP-programs were introduced in [10] to consider different ways in which the clients' specific information is used when computing answers; these proposed operators will temporally modify the public knowledge stored in the server just for obtaining the answer. Here, our research is not focussed in these contextual operators, and we will use the union operator $\cup$ as a simple context handling operator.

Below, in Fig. 1, the graphical representation of the client/server model proposed in [10], is shown depicting a client agent sending a contextual query, and the main components of a DeLP-server.

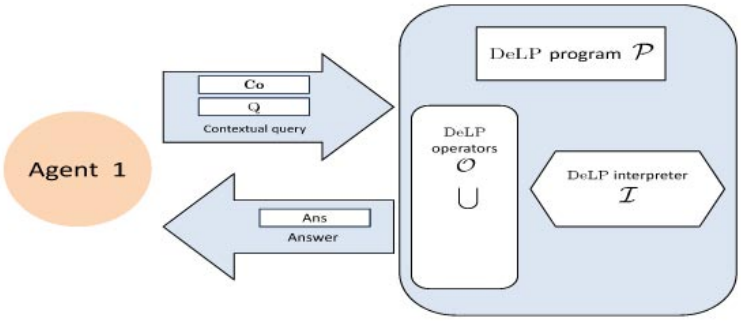


Fig. 1: Answer for a contextual query.

When reasoning with contradictory and dynamic information, the DeLP system builds arguments from the program. An argument $\mathcal{A}$ for a literal L is a minimal, non contradictory set of defeasible rules such that together with the programs' strict knowledge allows the derivation of L that will also be called the "conclusion" supported by $\mathcal{A}$, and for a given program the set of all possible arguments will be denoted as *Args*.

Given an argument $\mathcal{A}_1$ that is in conflict with other argument $\mathcal{A}_2$, to decide which one prevails, the two arguments must be compared using some criterion. Existing argumentation systems use a fixed comparison criterion embedded in the system.

3 Conditions based reasoning for the dynamic selection of criteria

As we have said, given two arguments $\mathcal{A}_1$ and $\mathcal{A}_2$ in conflict it is necessary to use a criterion to decide which argument prevails and if $\mathcal{A}_1$ is the prevailing argument, $\mathcal{A}_1$ is said to be a defeater of $\mathcal{A}_2$. If $\mathcal{A}_2$ is preferred to $\mathcal{A}_1$, then $\mathcal{A}_1$ will not defeat $\mathcal{A}_2$, and $\mathcal{A}_2$ prevails. If neither argument is preferred over the other, a blocking situation occurs and both arguments become defeated [9]. We will denote a preference criterion over arguments with C and a set of preference criteria $\mathcal{S} = \{\mathsf{C}_1, \mathsf{C}_2, \ldots, \mathsf{C}_n\}$. For our running example, we could assume the two criteria $\mathsf{C}_{comfort}$, that favors "comfort", and $\mathsf{C}_{security}$, that favors "security".

Definition 3 (Preference criterion) *Let Args be a set of arguments. A Preference criterion is a function* $\mathsf{C} : Args \times Args \longrightarrow \{\bot, \top\}$*, obtaining* $\top$ *when the first argument is preferred over the second, and* $\bot$ *otherwise.*

A contextual query includes the client's own information, and that information is used by the server to compute an answer. However, if in a client's query the server has to use a particular criterion, it will be necessary to change the structure of the contextual query to include them expanding it by including an expression that indicates to the server how to solve the query; to this end, a conditional expression will be added to the query expression.

A server will answer a query considering the preference criteria indicated by the client in a conditional-preference expression, or *cp-exp* for short; Definition 4 introduces the *cp-exps*. A *cp-exp* will be either a preference criterion, or *cp-exp* will be a guard $\mathcal{G}$ followed by two expressions $\mathcal{E}_1$ and $\mathcal{E}_2$ as explained below.

Definition 4 (Conditional-preference expression) *Given the criterion* $\mathsf{C} \in \mathcal{S}$ *and a set of literals* $\mathcal{G}$*. An expression* $\mathcal{E}$ *is a cp-exp iff:*

i) $\mathcal{E} = \mathsf{C}$*, or*

ii) $\mathcal{E} = [\mathcal{G} : \mathcal{E}_1; \mathcal{E}_2]$ *where* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are cp-exps.*

In case (i)*, the preference criterion* C *is applied, while in case* (ii) *the guard* $\mathcal{G}$ *is evaluated, if* $\mathcal{G}$ *verifies then* $\mathcal{E}_1$ *is applied, otherwise* $\mathcal{E}_2$ *is applied.*

As noted, a *cp-exp* may be associated to restrictive conditions. Let $\mathcal{P}'$ be a DeLP-program, to evaluate the set of literals $\mathcal{G}$ we will assume a function $eval(\mathcal{G}, \mathcal{P}')$ such that its range is $\{\top, \bot\}$, obtaining $\top$ iff for each literal $L \in \mathcal{G}$ there exist a strict derivation from $\mathcal{P}'$, *i.e.,* $\mathcal{P}' \vdash L$, $\bot$ otherwise.

We assume that the client agent may want that the criterion to be used be the one available by default already configured in the server. This criterion is denoted with the constant *"default"*. Therefore, if *cp-exp* $\mathcal{E} = \emptyset$ we assume that the used criterion will be the server default criterion. With a conditional-preference expression we are interested in providing a more declarative way of representing users preferences.

Example 3 *Consider the criteria* $\mathsf{C}_{comfort}$ *and* $\mathsf{C}_{security}$ *stated above. The criteria to establish whether a nearby lodging is recommended may be obtained by means of some of the following expressions:*

$$\mathcal{E}_1 = [\{downtown\} : \mathsf{C}_{comfort}, \mathsf{C}_{security}]$$
$$\mathcal{E}_2 = [\{dangerous\_area\} : [\{road\_blocked\} : \mathsf{C}_{security}, \mathsf{C}_{comfort}], default]$$

The expression $\mathcal{E}_1$ *represents the following conditional-preference expression: "If there is a strict derivation for downtown then use the comparison criterion preferring comfort, otherwise use the criterion favoring security". The expression* $\mathcal{E}_2$ *represents a nested condition: "If there is a strict derivation for dangerous_area then the expression* $[\{road\_blocked\} : \mathsf{C}_{security}, \mathsf{C}_{comfort}]$ *is evaluated, otherwise it uses the default criterion. The way of evaluating* $[\{road\_blocked\} : \mathsf{C}_{security}, \mathsf{C}_{comfort}]$ *is similar to the expression* $\mathcal{E}_1$*".*

The example above shows how simple expressions such as $\mathcal{E}_1$ and more complex expressions such as $\mathcal{E}_2$ could be built. As we will show in the following section, this example is of special interest since it will serve to show the different results two queries using these expressions will produce.

Thus, the client agents could indicate how their queries have to be solved by the server. For that reason, the *cp-exp* denoting the conditions the server has to consider to select a criterion, are included in the queries. This new type of contextual query will be called *conditional-preference based query.*

Definition 5 (Conditional-preference based query) *A conditional-preference based query* CQ *is a tuple* $[\mathcal{C}o, \mathcal{E}, Q]$*, where* $\mathcal{C}o$ *is a particular context for* CQ*,* $\mathcal{E}$ *is a cp-exp, and* Q *is a query.*

It is important to mention that *CQ* is an extension of the contextual query introduced in [10]. We refer the interested reader to [10] for details on those queries.

Example 4 *Going back to Example 2 and considering Example 3. Given the query "nearby_lodging", two conditional-preference based queries can be built:*

$$[\mathcal{P}_c, \mathcal{E}_1, nearby\_lodging]$$
$$[\mathcal{P}_c, \mathcal{E}_2, nearby\_lodging]$$

Given a *CQ*, we said that the set of criteria belonging to a *cp-exp* is the set of valid criteria, denoted $\mathbb{C}$, for that particular query. Following Example 4, the fact of having the same query but with distinct valid criteria set makes possible to vary the criterion used for solving each query.

As defined next, a DeLP-interpreter will be represented, in general, as a function such that given a program, a preference criterion and a query, it returns the corresponding answer.

Definition 6 (DeLP-interpreter) *Let* $\mathbb{P}$ *be the set of coherent DeLP-programs,* $\mathbb{C}$ *be a set of valid criteria and* $\mathbb{Q}$ *be the set of possible queries. A DeLP-interpreter is a function* $\mathcal{I} : \mathbb{P} \times \mathbb{C} \times \mathbb{Q} \longrightarrow \mathbb{R}$*, where* $\mathbb{R}$ *is the set of possible answers, i.e.,* $\mathbb{R} = \{$NO, YES, UNDECIDED, UNKNOWN$\}$.

As already mentioned, we propose a client/server interaction allowing the client agents to interact with a recommender server by sending conditional-preference based queries. Now, we formally present the concept of preference-based reasoning server.

Definition 7 (Conditional-preference based reasoning server) *A conditional-reference based reasoning server is a 4-tuple $\mathcal{CRS} = \langle \mathcal{I}, \mathcal{O}, \mathcal{P}, \mathcal{S} \rangle$, where $\mathcal{I}$ is a DeLP-interpreter, $\mathcal{O}$ is a set of DeLP-operators, $\mathcal{P}$ is a DeLP-program and $\mathcal{S}$ is a set of preference criteria.*

A $\mathcal{CRS}$ can accept queries from several clients and an agent can consult several servers. However, a $\mathcal{CRS}$ will answer only the queries that include criteria the server recognizes.

Consider a DeLP-program $\mathcal{P}$ modified with the context $\mathcal{C}o$, and a valid criterion $\mathsf{C} \in \mathbb{C}$. To evaluate a *cp-exp* $\mathcal{E}$ we will use a function $cond(\mathcal{E}, \mathcal{P})$ such that its range is the set of valid criteria $\mathbb{C}$, defined as follows.

Definition 8 (Condition Evaluation Function) *Let $\mathbb{E}$ be the set of all possible cp-exp, $\mathbb{P}$ be the set of coherent DeLP-programs, and $\mathbb{C}$ be a set of valid criteria, then we define the function cond with the following signature*

$$cond : \mathbb{E} \times \mathbb{P} \longrightarrow \mathbb{C},$$

and the evaluation of $\mathcal{E}$ in $\mathcal{P}$ is defined as:

i) $cond(\mathcal{E}, \mathcal{P}) = \mathsf{C}$ *if* $\mathcal{E} = \mathsf{C}$, *or*

ii) $cond(\mathcal{E}, \mathcal{P}) = cond(\mathcal{E}_1, \mathcal{P})$ *if* $\mathcal{E} = [\mathcal{G} : \mathcal{E}_1; \mathcal{E}_2]$ *and* $eval(\mathcal{G}, \mathcal{P}) = \top$, *or*

iii) $cond(\mathcal{E}, \mathcal{P}) = cond(\mathcal{E}_2, \mathcal{P})$ *if* $\mathcal{E} = [\mathcal{G} : \mathcal{E}_1; \mathcal{E}_2]$ *and* $eval(\mathcal{G}, \mathcal{P}) = \bot$.

Queries are answered using public knowledge stored in the server, plus individual knowledge sent with the query, and one of the criteria that a client agent sends as part of a conditional-preference expression. The answer will be obtained by means of an argumentative inference mechanism.

Definition 9 (Answer for a query) *Let $\mathcal{CRS} = \langle \mathcal{I}, \mathcal{O}, \mathcal{P}, \mathcal{S} \rangle$ be a conditional-preference based reasoning service, $PQ = [\mathcal{C}o, \mathcal{E}, Q]$ be a conditional-preference based query for $\mathcal{CRS}$, $\mathcal{P}'$ be a program modified with the context $\mathcal{C}o$, i.e., $\mathcal{P}' = \mathcal{P} \cup \mathcal{C}o$, and C_i be the criterion obtained from evaluating the expression $\mathcal{E}$. An answer for PQ from $\mathcal{CRS}$, denoted $Ans(\mathcal{CRS}, PQ)$, corresponds to the result of the function $\mathcal{I}(\mathcal{P}', \mathsf{C}_i, Q)$.*

4 Application example

In this section we will present a DeLP example showing how the answer to a query varies according to the criterion that results from evaluating a *cp-exp*. Let $\mathcal{P}_l$ and $\mathcal{P}_c$ be the DeLP-programs presented in Example 2, and consider the conditional-preference based queries from Example 4;

1. $[\mathcal{P}_c, \mathcal{E}_1, nearby\_lodging]$
2. $[\mathcal{P}_c, \mathcal{E}_2, nearby\_lodging]$

such that

$$\mathcal{E}_1 = [\{downtown\} : \mathsf{C}_{comfort}, \mathsf{C}_{security}]$$
$$\mathcal{E}_2 = [\{dangerous\_area\} : [\{road\_blocked\} : \mathsf{C}_{security}, \mathsf{C}_{comfort}], default]$$

In both queries, the same DeLP-program $\mathcal{P}' = \mathcal{P}_l \cup \mathcal{P}_c$ is obtained. From the program $\mathcal{P}'$ two arguments can be built: the argument $\mathcal{A}$ in favor of recommending a nearby lodging.

$$\mathcal{A} = \{\ nearby\_lodging \prec included\_AC.\ \}$$

and the argument $\mathcal{B}$ in favor of not recommending a nearby lodging:

$$\mathcal{B} = \{\ \sim nearby\_lodging \prec dangerous\_area.\ \}$$

Clearly, $\mathcal{A}$ and $\mathcal{B}$ are in conflict. To determine which one prevails, we have to establish the argument comparison criterion to be used. Consider the first conditional-preference based query presented above:

$$[\mathcal{P}_c, \mathcal{E}_1, nearby\_lodging]$$

Due to the strict derivation of "*downtown*" from $\mathcal{P}'$, we obtain the criterion $\mathsf{C}_{comfort}$ as the result of the function "*cond*". We assume that $\mathsf{C}_{comfort}$ establishes that $\mathcal{A}$ is preferred to $\mathcal{B}$, since the argument $\mathcal{A}$ has the information that the bedrooms in the lodge have air conditioning, then $\mathcal{A}$ will defeat to $\mathcal{B}$. In DeLP a query Q is warranted from a program $\mathcal{P}$ if there exists an undefeated argument $\mathcal{A}_1$ supporting Q. To establish whether the argument $\mathcal{A}$ is an undefeated argument, we will assume that the dialectical analysis proposed in [9] is performed. Since, this dialectical process assures that the conclusion *nearby_lodging* is warranted, then the answer for the query is YES.
Consider now the second conditional-preference based query presented above:

$$[\mathcal{P}_c, \mathcal{E}_2, nearby\_lodging]$$

after completing the whole argumentative process, the answer for *nearby_lodging* is NO, *i.e.,* the conclusion *nearby_lodging* is not warranted. In this case, the strict derivations of *dangerous_area* and *road_blocked* from $\mathcal{P}'$ are obtained, then the evaluation of $\mathcal{E}_2$ establish that the chosen criterion is $\mathsf{C}_{security}$. In order to obtain the answer for *nearby_lodging*, we assume that $\mathsf{C}_{security}$ determines that $\mathcal{B}$ is preferred to $\mathcal{A}$, since the argument $\mathcal{B}$ has the information that the vehicle is in a dangerous area. As $\mathcal{B}$ is a non-defeated argument, then the conclusion $\sim$*nearby_lodging* is warranted and the answer for the query is NO. For a detailed presentation of the dialectical analysis used for answer the two conditional-preference based queries introduced in this paper see [9].

As mentioned, in [10] the proposed server is configured to use a fixed comparison criterion embedded in the system. Thus, the answers to our example queries will be always solved using the same criterion. In contrast to [10], in the complete example we show that with our approach the same query with the same context but with different conditional-preference expressions can give different criteria, and possibly different answers. This was one of our goals.

5 Conclusions, related and future work

We presented a model that allows an argumentative reasoning server to dynamically select the argument comparison criterion to be used. For this, we formally defined the notion of conditional-preference expression, which is part of a new type of contextual query, called conditional-preference based query. We showed how these expressions are evaluated by means of a function, called "*cond*", which determines which criterion prevails for the query. DeLP was proposed as the knowledge representation and reasoning system, therefore the DeLP-interpreter is in charge of solving the queries. In Section 4 an example was presented where an agent performs two queries with the same context but with different conditional-preference expressions, getting different results. We showed how in the proposed model, argument comparison criteria are directly related to the inferences obtained by an agent.

Our approach was in part inspired by [10], where several servers can be created, and knowledge can be shared through them. Nevertheless, in contrast with us, they use a preference criteria embedded into the interpreter, *i.e.,* to answer a query, the server is configured to use the same specific criterion. In fact, we provide clients with the possibility of indicating to the server what criteria could use to compute the answer for a specific query.

In [4], an approach where the preference is subjected to conditional dependence was proposed. A preference relation is defined as a total pre-order (a ranking) over some set of variables such that the preference over the values of one variable depends on the value of others. Their main contribution is a graphical representation of preferences that reflects conditional dependence and independence of preference statements under a *ceteris paribus* (all else being equal) interpretation. Similar to us, the authors present a model for representing and reasoning with the user preferences, where conditional preference expressions are permitted. In contrast with us, they provide a framework where the preferences are considered for decision making where the space of possible actions or decisions available to someone is fixed, with well-understood dynamics, conversely, in our framework the situation is different, *i.e.,* the selected application domains are dynamic and agents deal with incomplete and contradictory information; for that reason, our research is focused on argumentative systems.

As future work we are developing an implementation of a DeLP-server that can dynamically handle conditional based preference criteria. We are also studying the possibility of developing more powerful comparison criteria expressions

that allow more expressive combinations. Another extension will be to integrate our proposed framework with others argumentative systems similar to DeLP.

Acknowledgements: This work is partially supported by CONICET, Universidad Nacional de Entre Ríos (PID-UNER 7041), Universidad Nacional del Sur, SGCyT.

References

1. Alsinet, T., Chesñevar, C.I., Godo, L., Simari, G.R.: A logic programming framework for possibilistic argumentation: Formalization and logical properties. Fuzzy Sets and Systems 159(10), 1208–1228 (2008)
2. Amgoud, L., Parsons, S.: Agent dialogues with conflicting preferences. In: ATAL. pp. 190–205 (2001)
3. Antoniou, G., Maher, M.J., Billington, D.: Defeasible logic versus logic programming without negation as failure. J. Log. Program. 42(1), 47–57 (2000)
4. Boutilier, C., Brafman, R.I., Hoos, H.H., Poole, D.: Reasoning with conditional ceteris paribus preference statements. In: Laskey, K.B., Prade, H. (eds.) UAI. pp. 71–80. Morgan Kaufmann (1999)
5. Boutilier, C.: Toward a logic for qualitative decision theory. In: KR. pp. 75–86 (1994)
6. Capobianco, M., Chesñevar, C.I., Simari, G.R.: Argumentation and the dynamics of warranted beliefs in changing environments. Autonomous Agents and Multi-Agent Systems 11(2), 127–151 (2005)
7. Capobianco, M., Simari, G.R.: A proposal for making argumentation computationally capable of handling large repositories of uncertain data. In: SUM. pp. 95–110 (2009)
8. Deagustini, C.A.D., Fulladoza Dalibón, S.E., Gottifredi, S., Falappa, M.A., Chesñevar, C.I., Simari, G.R.: Relational databases as a massive information source for defeasible argumentation. Knowledge-Based Systems (to appear) (2013)
9. García, A.J., Simari, G.R.: Defeasible logic programming: An argumentative approach. Theory and Practice of Logic Programming (TPLP) 4, 95–138 (2004)
10. García, A.J., Rotstein, N.D., Tucat, M., Simari, G.R.: An argumentative reasoning service for deliberative agents. In: KSEM. pp. 128–139 (2007)
11. Konstan, J.A.: Introduction to recommender systems: Algorithms and evaluation. ACM Trans. Inf. Syst. 22(1), 1–4 (2004)
12. Li, M., Vo, Q.B., Kowalczyk, R.: Majority-rule-based preference aggregation on multi-attribute domains with cp-nets. In: AAMAS. pp. 659–666 (2011)
13. Loui, R.P.: Defeat among arguments: a system of defeasible inference. Computational Intelligence 3, 100–106 (1987)
14. Maher, M.J., Rock, A., Antoniou, G., Billington, D., Miller, T.: Efficient defeasible reasoning systems. International Journal on Artificial Intelligence Tools 10(4), 483–501 (2001)
15. Resnick, P., Varian, H.R.: Recommender systems - introduction to the special section. Commun. ACM 40(3), 56–58 (1997)
16. Simari, G.R., Loui, R.P.: A mathematical treatment of defeasible reasoning and its implementation. Artificial Intelligence 53(2-3), 125–157 (1992)
17. Stolzenburg, F., García, A.J., Chesñevar, C.I., Simari, G.R.: Computing generalized specificity. J. of Applied Non-Classical Logics 13(1), 87–113 (2003)
18. Vreeswijk, G.: Abstract argumentation systems. Artificial Intelligence 90(1-2), 225–279 (1997)

On semantics in dynamic argumentation frameworks

Maria Laura Cobo, Diego C. Martinez, and Guillermo R. Simari

Artificial Intelligence Research and Development Laboratory (LIDIA)
Department of Computer Science and Engineering, Universidad Nacional del Sur
Av. Alem 1253 - (8000) Bahía Blanca - Bs. As. - Argentina
{mlc,dcm,grs}@cs.uns.edu.ar
http://www.cs.uns.edu.ar/lidia

Abstract. A Timed Abstract Argumentation Framework is a novel formalism where arguments are only valid for consideration in a given period of time, which is defined for every individual argument. Thus, the attainability of attacks and defenses is related to time, and the outcome of the framework may vary accordingly. In this work we study the notion of stable extensions applied to timed-arguments. The framework is extended to include intermittent arguments, which are available with some repeated interruptions in time.

Keywords: Timed Abstract Argumentation, Abstract Argumentation, Timed Information

1 Introduction

One of the main concerns in Argumentation Theory is the search for rationally based positions of acceptance in a given scenario of arguments and their relationships. This task requires some level of abstraction in order to study pure semantic notions. Abstract argumentation systems [11, 16, 2, 3] are formalisms for argumentation where some components remain unspecified, being the structure of an argument the main abstraction. In this kind of system, the emphasis is put on the semantic notion of finding the set of accepted arguments. Most of these systems are based on the single abstract concept of *attack* represented as an abstract relation, and extensions are defined as sets of possibly accepted arguments. For two arguments $\mathcal{A}$ and $\mathcal{B}$, if $(\mathcal{A}, \mathcal{B})$ is in the attack relation, then the acceptance of $\mathcal{B}$ is conditioned by the acceptance of $\mathcal{A}$, but not the other way around. It is said that argument $\mathcal{A}$ *attacks* $\mathcal{B}$, and it implies a priority between conflicting arguments.

The simplest abstract framework is defined by Dung in [11]. It only includes a set of abstract arguments and a binary relation of attack between arguments. Several semantics notions are defined and the Dung's argument extensions became the foundation of further research. Other proposals extends Dung's framework by the addition of new elements, such as preferences between arguments [2, 7] or subarguments [14]. Other authors use the original framework to elaborate new extensions [12, 5]. All of these proposals are based on varied abstract formalizations of arguments and attacks.

In this scenario, the combination of time and argumentation is a novel research line. In [13] a calculus for representing temporal knowledge is proposed, and defined in terms of propositional logic. This calculus is then considered with respect to argumentation,

where an argument is defined in the standard way: an argument is a pair constituted by a minimally consistent subset of a database entailing its conclusion. This work is thus related to [4].

In [9, 10] a novel framework is proposed, called *Timed Abstract Framework* (TAF), combining arguments and temporal notions. In this formalism, arguments are relevant only in a period of time, called its *availability interval*. This framework mantains a high abstract level in an effort to capture intuitions related with the dynamic interplay of arguments as they become available and cease to be so. The notion of *availability interval* refers to an interval of time in which the argument can be legally used for the particular purpose of an argumentation process. Thus, this kind of timed-argument has a limited influence in the system, given by the temporal context in which these arguments are taken into account.

Timed abstract frameworks capture the previous argument model by assigning arguments to an availability interval of time. In [10] a skeptical, timed interval-based semantics is proposed, using admissibility notions. As arguments may get attacked during a certain period of time, defense is also time-dependant, requiring a proper adaptation of classical acceptability. In [9], algorithms for the characterization of defenses between timed arguments are presented.

In [9] a natural expansion of timed argumentation frameworks by considering arguments with more than one availability interval is introduced. These arguments are called *intermittent arguments*, available with (possibly) some repeated interruptions in time. In all of these scenarios arguments may become relevant, or cease to be so, depending on time-related factors.

This paper is organized as follows. In the next section we recall time representation notions, where time-intervals are presented. Thereafter, the terminology used in this work are defined, towards the presentation of our Timed Abstract Argumentation Framework with intermittent arguments in Section 3. The notion of stable extension is presented in Section 4. The relation among steadiness and dynamics is analyzed in Section 5. Finally, conclusions and future work are discussed.

2 Time representation

In order to capture a time-based model of argumentation, we enrich the classical abstract frameworks with temporal information regarding arguments. The problem of representing temporal knowledge and temporal reasoning arises in a lot of disciplines, including Artificial Intelligence. There are many ways of representing temporal knowledge. A usual way to do this is to determine a *primitive* to represent time, and its corresponding *metric relations* [1, 15]. In this work we will use *temporal intervals of discrete time* as primitives for time representation, and thus only metric relations for intervals are applied.

Definition 1 *[Temporal Interval] An interval is a pair build from* $a, b \in \mathbb{Z} \cup \{-\infty, \infty\}$*, in one of the following ways:*

- $[a, a]$ *denotes a set of time moments formed only by moment* a*.*
- $[a, \infty)$ *denotes a set of moments formed by all the numbers in* $\mathbb{Z}$ *since* a *(including* a*).*

- $(-\infty, b]$ *denotes a set of moments formed by all the numbers in* $\mathbb{Z}$ *until moment* i *(including* b*).*
- $[a, b]$ *denotes a set of moments formed by all the numbers in* $\mathbb{Z}$ *from moment* i *until moment* j *(including both* a *and* b*).*
- $(-\infty, \infty)$ *a set of moments formed by all the numbers in* $\mathbb{Z}$*.*

The moments a, b *are called endpoints. The set of all the intervals defined over* $\mathbb{Z} \cup \{-\infty, \infty\}$ *is denoted* Υ*.*

For example, $[5, 12]$ and $[1, 200]$ are intervals. If X is an interval then X^-, X^+ are the corresponding endpoints ($i.e.$, $X = [X^-, X^+]$). An endpoint may be a point of discrete time, identified by an integer number, or infinite.

We will usually work with sets of intervals (as they will be somehow related to arguments). Thus, we introduce several definitions and properties needed for semantic elaborations.

In the following section we present Timed Abstract Argumentation Frameworks with intermittent arguments.

3 Timed Argumentation Framework

As remarked before, in Timed Argumentation Frameworks [9] the consideration of time restrictions for arguments is formalized through an *availability function*, which defines a temporal interval for each argument in the framework. This interval states the period of time in which an argument is available for consideration in the argumentation scenario. The formal definition of our timed abstract argumentation framework follows.

Definition 2 *A timed abstract argumentation framework (TAF) is a 3-tuple* $\langle Args, Atts, \mathcal{A}v \rangle$ *where* $Args$ *is a set of arguments,* $Atts$ *is a binary relation defined over* $Args$ *and* $\mathcal{A}v$ *is the availability function for timed arguments, defined as* $\mathcal{A}v : Args \rightarrow \wp(\Upsilon)$*.*

Example 1 *The triplet* $\langle Args, Atts, \mathcal{A}v \rangle$*, where* $Args = \{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}\}$*,* $Atts = \{(\mathcal{B}, \mathcal{A}), (\mathcal{C}, \mathcal{B}), (\mathcal{D}, \mathcal{A}), (\mathcal{E}, \mathcal{D})\}$ *and the availability function is defined as*

$Args$	$\mathcal{A}v$	$Args$	$\mathcal{A}v$
$\mathcal{A}$	$\{[10, 40], [60, 75]\}$	$\mathcal{B}$	$\{[30, 50]\}$
$\mathcal{C}$	$\{[20, 40], [45, 55], [60, 70]\}$	$\mathcal{D}$	$\{[47, 65]\}$
$\mathcal{E}$	$\{(-\infty, 44]\}$		

is a timed abstract argumentation framework.

The framework of Example 1 can be depicted as in Figure 1, using a digraph where nodes are arguments and arcs are attack relations. An arc from argument $\mathcal{X}$ to argument $\mathcal{Y}$ exists if $(\mathcal{X}, \mathcal{Y}) \in Atts$. Figure 1 also shows the time availability of every argument, as a graphical reference of the $\mathcal{A}v$ function. It is basically the framework's evolution in time. Endpoints are marked with a vertical line, except for $-\infty$ and ∞. For space reasons, only some relevant time points are numbered in the figure. As stated before, the availability of arguments is tied to a temporal restriction. Thus, an attack to an

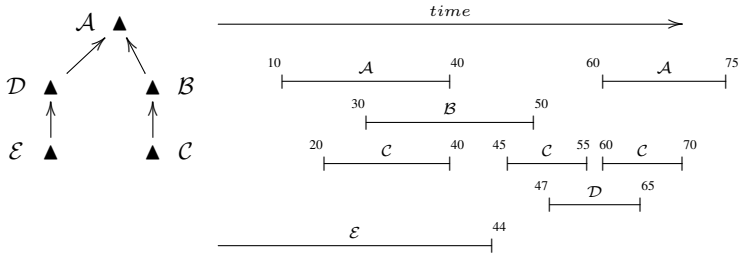


Fig. 1. Framework of Example 1

argument may actually occur only if both the attacker and the attacked argument are available. In other words, an attack between two arguments may be *attainable*, under certain conditions. *Attainable attacks* are attacks that will eventually occur in some period of time. In order to formalize this, we need to compare time intervals, using the previously defined metric relations.

Definition 3 *Let $\Phi = \langle Args, Atts, \mathcal{A}v\rangle$ be a TAF, and let $\{\mathcal{A}, \mathcal{B}\} \subseteq Args$ such that $(\mathcal{B}, \mathcal{A}) \in Atts$. The attack $(\mathcal{B}, \mathcal{A})$ is said to be attainable if $I_{\mathcal{A}}$ overlaps $I_{\mathcal{B}}$, for some $I_{\mathcal{A}} \in \mathcal{A}v(\mathcal{A})$ and $I_{\mathcal{B}} \in \mathcal{A}v(\mathcal{B})$. The attack is said to be attainable in $\mathcal{A}v(\mathcal{A}) \Cap \mathcal{A}v(\mathcal{B})$. The set of intervals where an attack $(\mathcal{B}, \mathcal{A})$ is attainable will be noted as IntSet$((\mathcal{B}, \mathcal{A}))$*

Note that an attack is attainable if the availability of both the attacker and the attacked argument overlaps in at least one moment of time.

Example 2 *Consider the timed argumentation framework of Example 1. The attacks $(\mathcal{D}, \mathcal{A})$ and $(\mathcal{B}, \mathcal{A})$ are both attainable in the framework. Attack $(\mathcal{D}, \mathcal{A})$ is attainable since $[47, 65] \cap [60, 75]$ is non-empty, with $[47, 65] \in \mathcal{A}v(\mathcal{D})$ and $[60, 75] \in \mathcal{A}v(\mathcal{A})$. Attack $(\mathcal{B}, \mathcal{A})$ is attainable since $[30, 50] \cap [10, 40]$, is $[30, 40]$. Recall that $[30, 50] \in \mathcal{A}v(\mathcal{B})$, $[10, 40] \in \mathcal{A}v(\mathcal{A})$. The attack $(\mathcal{C}, \mathcal{B})$ is also attainable. Since $\mathcal{A}v(\mathcal{C}) = \{[20, 40], [30, 50]\}$ and $\mathcal{A}v(\mathcal{B}) = \{[30, 50]\}$ then we can assure the attainability of the attack because $\mathcal{A}v(\mathcal{C}) \Cap \mathcal{A}v(\mathcal{B})$ is non-empty. The attack is then attainable at $\{[30, 40], [45, 50]\}$, $i.e.$ in $\mathcal{A}v(\mathcal{C}) \Cap \mathcal{A}v(\mathcal{B})$. The attack $(\mathcal{E}, \mathcal{D})$ is not attainable, since the intersection among $(-\infty, 45]$ and $[47, 65]$ is empty. The arguments involved in this attack are never available at the same time.*

The set of all the attainable attacks in the framework Φ is denoted $AttAtts_{\Phi}$. It is also possible to define the attainability of attacks at a particular timed intervals, as shown next.

Definition 4 *Let $\Phi = \langle Args, Atts, \mathcal{A}v\rangle$ be a TAF, and let $\{\mathcal{A}, \mathcal{B}\} \subseteq Args$ such that $(\mathcal{B}, \mathcal{A}) \in Atts$. The attack $(\mathcal{B}, \mathcal{A})$ is said to be attainable at I if: $I \cap \mathcal{A}v(\mathcal{A}) \neq [\,]$ and the following condition holds: $I \cap I_{\mathcal{A}}$ overlaps $I_{\mathcal{B}}$, for some $I_{\mathcal{A}} \in \mathcal{A}v(\mathcal{A})$ and $I_{\mathcal{B}} \in \mathcal{A}v(\mathcal{B})$.*

The set of attainable attacks of Φ at interval I is denoted $AttAtts^{I}_{\Phi}$.

Example 3 *Consider the timed argumentation framework of Example 1. The set* $AttAtts_{\Phi}$ *is:* $\{(\mathcal{D},\mathcal{A}),(\mathcal{B},\mathcal{A}),(\mathcal{C},\mathcal{B})\}$. *The set* $AttAtts_{\Phi}^{[35,40]}$ *is* $\{(\mathcal{B},\mathcal{A}),(\mathcal{C},\mathcal{B})\}$. *The attack* $(\mathcal{D},\mathcal{A})$ *is in* $AttAtts_{\Phi}$ *but it is not in* $AttAtts_{\Phi}^{[35,40]}$, *since* $[35,40] \cap [47,65]$ *is the emptyset. The attack* $(\mathcal{B},\mathcal{A})$ *is in* $AttAtts_{\Phi}$ *and is also in* $AttAtts_{\Phi}^{[35,40]}$, *since* $[35,40] \cap [10,40] = [35,40]$ *and* $[35,40] \cap [30,50] = [35,40]$. *Note that* $[10,40] \in \mathcal{A}v(\mathcal{A})$ *and* $[30,50] \in \mathcal{A}v(\mathcal{B})$.

The definition of attainability of attacks can be attached to particular time points too. The set of attainable attacks of Φ at moment i is denoted $AttAtts_{\Phi}(i)$ and is defined as $AttAtts_{\Phi}(i) = AttAtts_{\Phi}^{[i,i]}$.

4 Semantics for Timed Argumentation

In [9, 10, 8] several semantic notions for timed frameworks are introduced. Admissibility semantics are captured by considering temporal defense. As attacks may occur only on a period of time (that in which the participants are available), argument defense is also occasional. In [10] a skeptical, timed interval-based semantics is proposed, using admissibility notions. The classical definition of acceptability is adapted to a timed context. The complexity of this adaptation lies on the fact that defenses may occur sporadically and hence the focus is put on finding *when* the defense takes place. For example, an argument $\mathcal{A}$ may be defended by $\mathcal{X}$ in the first half of an availability interval, and later by an argument $\mathcal{Y}$ in the second half. Although $\mathcal{X}$ is not capable of providing a full defense, argument $\mathcal{A}$ is defended while $\mathcal{A}$ is available. In other words, defenders *take turns* to provide a defense.

In [8] a notion of stable extension is introduced which considers the global evolution of a timed framework. This requires the definition of the notion of t-profile: a pair formed by an argument and a set of intervals in which this argument is considered. Since arguments are related to time, a t-profile of an argument $\mathcal{X}$ is the formal reference of $\mathcal{X}$ within several frames of time, which are subintervals of the original availability intervals of $\mathcal{X}$. Hence, an argument is not considered stand-alone in a specific moment of time, but associated with a set of intervals. A t-profile attacks another t-profile if an attack is formally defined between its arguments and at least one interval of time of each profile is overlapping. A set of t-profiles is a collection of arguments which are considered within different intervals of time, not necessarily overlapping. The timed notion of *stable extension* is later defined, not as a set of arguments but as a set S of t-profiles such that every t-profile denotes intervals of time in which a given argument attacks other available arguments. Hence, the arguments in these t-profiles may collectively form a stable set. Notoriously, an argument $\mathcal{X}$ may appear in t-profiles *inside* and *outside* this timed stable set simultaneously, but with different intervals of time since an argument may become attacked or not as time evolves. Thus, an argument may gets in and out a stable set depending on time, but in any point of time a set of arguments is characterized which attacks any other argument not included in that set.

Beyond the previous time-based semantics, since at any moment in the evolution of the timed framework there may be active arguments with available attacks, it is possible

to apply Dung's classical semantics at any timepoint. When arguments become available or cease to be so, these semantic consequences may change. This is addressed in the following section.

5 Steadiness in dynamic argumentation

A timed argumentation framework is a natural model for argumentation dynamics, where the set of arguments is not fixed and evolves through time, $i.e.$ arguments may appear or disappear from the framework. These evolution may cause changes in the semantic consequences of the overall set of arguments and how an argument impacts on the outcome of the argumentation framework depends naturally on the particular semantics. We are mainly interested in the study of periods of time in which some semantic properties are unaffected by this argument dynamics. We will refer to these as *steady* intervals of a timed framework. For instance, the introduction of a new argument may not change any argument extension of a given semantics, as shown in the following example.

Example 4 *Consider the TAF of Figure 2. Argument $\mathcal{A}$ is attacked by arguments $\mathcal{B}$ and $\mathcal{C}$, although in different moments in time. However, argument $\mathcal{D}$ provides a defense for $\mathcal{A}$ whenever it is attacked. Hence, although an argument cease to exist and another one begins, $\{\mathcal{A}, \mathcal{D}\}$ is an admissible set in* $[10, 30]$.

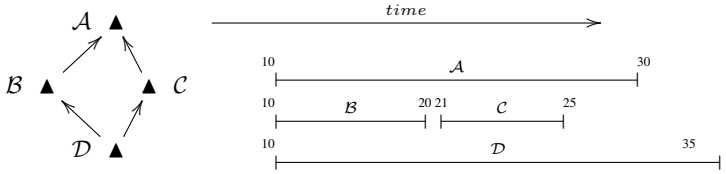


Fig. 2. Steady notions

There is also another view of steady intervals when conclusions of arguments are taken into account. Since arguments support conclusions, then this conclusions are kept in time. However, a new argument may change argument extensions while preserving the set of conclusions supported by those arguments.

Example 5 *Consider the TAF of Figure 3. Argument $\mathcal{A}$ supporting conclusion h is free of attackers in* $[10, 20]$. *Later on, it is attacked by argument $\mathcal{C}$ supporting conclusion g in* $[20, 40]$. *However, conclusion h is also supported by argument $\mathcal{B}$ which is not attacked by $\mathcal{C}$ and although $\mathcal{A}$ lacks of defenders, conclusion h is sustained in* $[10, 50]$.

The following definitions provide basic notions for the study of the evolution of a timed framework.

Definition 5 *Let $\mathcal{A}$ be an argument and let I be an interval.*

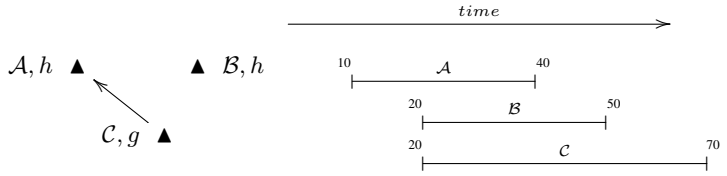


Fig. 3. Steady notions

- $Args(i)$ *is the set of arguments available at a given timepoint* i*.*
- *An argument* $\mathcal{A}$ *is said to* occur *in* I *if* $\mathcal{A} \in Args(k)$ *for some* $k \in I$*.*
- *An argument* $\mathcal{A}$ *is said to* fully occur *in* I *if* $\mathcal{A} \in Args(k)$ *for all* $k \in I$*.*
- *An argument* $\mathcal{A}$ *is said to* partially occur *in* I *if* $\exists k \in I$ *such that* $\mathcal{A} \notin Args(k)$*.*

Definition 6 *Two intervals* I_1 *and* I_2 *are consecutive if* $I_1^+ = I_2^- - 1$*.*

In a timed argumentation framework, arguments may be available or cease to be so as times goes by. An interval in which no changes occur in the framework is said to be *static*, as defined next.

Definition 7 *A static interval for a TAF is a period of time* $I = [i, j]$ *such that* $\forall k, m \in I, Args(k) = Args(m)$*. A maximal static interval is a static interval not included in another static interval.*

A static interval is the first notion of steadiness in a timed framework. In the framework of Figure 2 the intervals $[10, 15]$,$[10, 20]$ $[21, 25]$ and $[26, 35]$ are all static intervals.

Definition 8 *Let* I_1, I_2 *be two consecutive maximal static intervals. The pair* (I_1^+, I_2^-) *is said to be the changing leap of* I_1 *to* I_2

Note that the changing leap denotes a transition, since something has ocurred that breaks static periods of time. This is stated in the following proposition.

Proposition 1 *Let* I_1 *and* I_2 *be two consecutive maximal static intervals. Then* $Args(I_1^+) \neq Args(I_2^-)$*.*

What is really interesting about changing leaps is the ability to affect semantic consequences. For instance, a single new argument may cause several arguments to be dropped out of argument extensions. The following definition, inspired from [6], characterizes the set of all the argument extensions induced by a given semantic.

Definition 9 *Let* $\mathcal{S}$ *be an argumentation semantics. The set* $\mathcal{E}_{\mathcal{S}}(i)$ *is the set of all the extensions under semantic* $\mathcal{S}$ *at timepoint* i*.*

In the timed framework of Figure 2, given $\mathcal{S} = admissibility$, then $\mathcal{E}_{\mathcal{S}}(15) = \{\{\mathcal{A}, \mathcal{D}\}, \{\mathcal{D}\}\}$ and $\mathcal{E}_{\mathcal{S}}(32) = \{\{\mathcal{D}\}\}$.

Definition 9 leads to a semantic notion of steady intervals: those in which the set of extensions induced by a given semantics does not change over time. Thus, in every timepoint of the interval the set of extensions is the same. This is formalized in the following definition.

Definition 10 *Let $\mathcal{S}$ be an argumentation semantic for TAF. A steady interval for $\mathcal{S}$ is a period of time $I = [i, j]$ such that $\forall k, m \in I, \mathcal{E}_{\mathcal{S}}(k) = \mathcal{E}_{\mathcal{S}}(m)$. A maximal steady interval for $\mathcal{S}$ is a steady interval not included in another steady interval. The set of all the extensions during a steady interval I is denoted $\mathcal{E}_{\mathcal{S}}^{\mathcal{I}}$.*

In this kind of intervals the framework is *steady* since it is not semantically disturbed by changes in the set of arguments, if any. In the timed framework of Figure 2, the intervals $I_1 = [10, 30]$ and $I_2 = [31, 35]$ are steady intervals for admissibility.

Definition 11 *Let I_1 and I_2 be two maximal steady intervals for semantics $\mathcal{S}$. The changing leap of I_1 and I_2 is said to be a semantic leap of $\mathcal{S}$. A changing leap that is not a semantic leap is said to be irrelevant to $\mathcal{S}$.*

Not every change in the set of arguments is a semantic leap. In the timed framework of Example 2 argument $\mathcal{B}$ is *replaced* by argument $\mathcal{C}$, but this situation does not affect the admissible set $\{\mathcal{A}, \mathcal{D}\}$. However, if no change occurs, then naturally the semantic consequences remain unchanged, as stated in the following proposition.

Proposition 2 *Every static interval is a steady interval.*

A semantic leap is interesting since it denotes a changing leap with an impact in the outcome of the framework. It is interesting to identify arguments introduced and discarded by the semantic leap.

Definition 12 *Let I_1, I_2 be two consecutive steady intervals. The sets $AV_{in}(I_1, I_2)$ and $AV_{out}(I_1, I_2)$ are defined as*

- $AV_{in}(I_1, I_2) = Args(I_2^-) \setminus Args(I_1^+)$
- $AV_{out}(I_1, I_2) = Args(I_2^+) \setminus Args(I_1^-)$

Although semantic extensions may change between two consecutive steady intervals, an argument may still be included in some extension. Since it spans between two interval, this argument is not added nor deleted in the semantic leap.

Definition 13 *Let Φ be a TAF. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two argumentation semantics. Let $sl_{\mathcal{S}_1}^{i,j}$ and $sl_{\mathcal{S}_2}^{i,j}$ be the set of all the semantic leaps of Φ in $[i, j]$ for semantics $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. Two argumentation semantics $\mathcal{S}_1$ and $\mathcal{S}_2$ are said to be chained in $[i, j]$ if $sl_{\mathcal{S}_1}^{i,j} = sl_{\mathcal{S}_2}^{i,j}$*

Chained semantics share semantic leaps in a given interval, although the outcome of these semantics may differ. Semantic leaps keep track of what information does not change with the evolution of the framework.

Definition 14 *An argumentation framework is said to be well-formed in interval I if it is well-formed in any timepoint of I, $i.e.$ it is cycle-free in that timepoint.*

Proposition 3 *If an argumentation framework is well-formed at interval I, then the grounded and stable semantics are chained in I.*

Proof: As stated in [11], for well-formed argumentation frameworks the grounded and stable semantics coincide. If an argumentation framework is well-formed in interval I, then any changing leap does not introduce cycles. Hence, any changing leap causing a semantic change in the grounded extension will cause a change in the stable extension and viceversa.□

Proposition 4 *Let Φ be a TAF. Let α be a changing leap. If $\alpha = (i, i+1)$ is an irrelevant leap for admissiblity, but a semantic leap of stable, then Φ is not well-formed at timepoint i.*

Proof: If α is irrelevant for admissibility, then no admissible extension is changed after α. It means that all the arguments in an admissible extension before α keep their defenders after α. Thus α (a) removes an attacked attacker or (b) introduces new attacking arguments which are attacked in turn by arguments in a previous admissible set. Suppose Φ is well-formed after α. Then case (b) does not introduces new argument cycles (the same is trivially true for case (a)). Since a new argument $\mathcal{A}$ is not introducing cycles, then Since α is a semantic leap for stable semantics, then $\mathcal{E}_{stable}(i) \neq \mathcal{E}_{stable}(i+1)$.

Definition 15 *Let $\mathcal{A}$ be an argument. The supportive interval of argument $\mathcal{A}$ is a maximal interval $I = [i, j]$ such that for any timepoint m in I, it holds that $\mathcal{A} \in E$ for some $E \in \mathcal{E}_{\mathcal{S}}(m)$.*

A supportive interval for an argument can span several steady intervals and they share endpoints with steady intervals.

Remark 1 *It is possible for two consecutive steady intervals have the same set of warranted arguments, although in different extensions.*

Proposition 5 *Let $I = [i, j]$ be a steady interval and let $\mathcal{A}$ be an argument in some extension E from $\mathcal{E}^{\mathcal{I}}_{\mathcal{S}}$. Then I is a subinterval of a supportive interval of $\mathcal{A}$.*

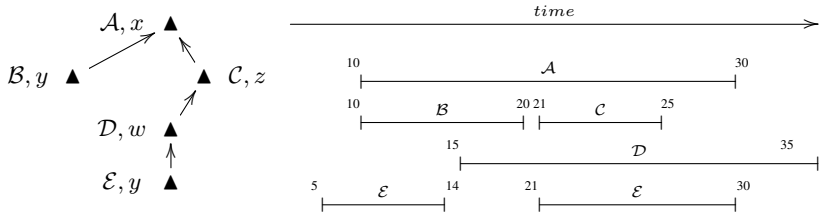


Fig. 4. Steady notions

As shown in Example 5 and in framework depicted on Figure 4 there is another notion of steadiness in timed argumentation frameworks, and it is related to the fact that several arguments may support the same conclusion. In the framework depicted on Figure 4 y is supported on $[5, 30]$ although in each particular moment of the interval y is granted through argument $\mathcal{B}$, $\mathcal{D}$ or both. The formalization of this ideas is currently being explored.

6 Conclusions and future work

In this work we presented an analysis over timed argumentation frameworks. The main idea was to find stable time periods, periods of time where extensions do not change. This periods are interesting in frameworks dynamics, since future changes only affects some of these periods instead of affecting the whole framework. Concepts and relations related with this notion where presented. Future work has several directions. Some of them are a deeper analysis of steadiness concept properties; other forms of steadiness are of particular interest. Steadiness of argument's conclusions is currently being analyzed.

References

1. Allen, J.: Maintaining knowledge about temporal intervals. Communications of the ACM (26), 832–843 (1983)
2. Amgoud, L., Cayrol, C.: On the acceptability of arguments in preference-based argumentation. In: 14th Conference on Uncertainty in Artificial Intelligence (UAI'98). pp. 1–7. Morgan Kaufmann (1998)
3. Amgoud, L., Cayrol, C.: A reasoning model based on the production of acceptable arguments. In: Annals of Mathematics and Artificial Intelligence, vol. 34, 1-3, pp. 197–215 (2002)
4. Augusto, J.C., Simari, G.R.: Temporal defeasible reasoning. Knowl. Inf. Syst. 3(3), 287–318 (2001)
5. Baroni, P., Giacomin, M.: Resolution-based argumentation semantics. In: Proc. of 2nd International Conf. on Computational Models of Argument (COMMA 2008). pp. 25–36 (2008)
6. Baroni, P., Giacomin, M.: On principle-based evaluation of extension-based argumentation semantics. Artif. Intell. 171(10-15), 675–700 (2007)
7. Bench-Capon, T.: Value-based argumentation frameworks. In: Proc. of Nonmonotonic Reasoning. pp. 444–453 (2002)
8. Cobo, M.L., Martínez, D.C., Simari, G.R.: Stable extensions in timed argumentation frameworks. In: Modgil, S., Oren, N., Toni, F. (eds.) TAFA. Lecture Notes in Computer Science, vol. 7132, pp. 181–196. Springer (2011)
9. Cobo, M., Martinez, D., Simari, G.: An approach to timed abstract argumentation. In: Proc. of Int. Workshop of Non-monotonic Reasoning 2010 (2010)
10. Cobo, M., Martinez, D., Simari, G.: On admissibility in timed abstract argumentation frameworks. In: Coelho, H., Studer, R., Wooldridge, M. (eds.) ECAI. Frontiers in Artificial Intelligence and Applications, vol. 215, pp. 1007–1008. IOS Press (2010)
11. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77(2), 321–358
12. Jakobovits, H.: Robust semantics for argumentation frameworks. Journal of Logic and Computation 9(2), 215–261 (1999)
13. Mann, N., Hunter, A.: Argumentation using temporal knowledge. In: Proc. of 2nd International Conf. on Computational Models of Argument (COMMA 2008). pp. 204–215 (2008)
14. Martínez, D.C., García, A.J., Simari, G.R.: Modelling well-structured argumentation lines. In: Proc. of XX IJCAI-2007. pp. 465–470 (2007)
15. Meiri, I.: Combining qualitative and quantitative contraints in temporal reasoning. In: Proceedings of AAAI '92. pp. 260–267 (1992)
16. Vreeswijk, G.A.W.: Abstract argumentation systems. Artificial Intelligence 90(1–2), 225–279 (1997)

Multi-criteria Argumentation-Based Decision Making within a BDI Agent

Cecilia Sosa Toranzo, Marcelo Errecalde, and Edgardo Ferretti

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Universidad Nacional de San Luis, Ejército de los Andes 950, San Luis - Argentina
e-mails:{csosatoranzo, merreca, ferretti}@unsl.edu.ar

Abstract. The BDI model, as a practical reasoning architecture aims at making decisions about what to do based on cognitives notions as beliefs, desires and intentions. However, during the decision making process, BDI agents also have to make background decisions like choosing what intention to achieve next from a set of possibly conflicting desires; which plan to execute from among the plans that satisfy a given intention; and whether is necessary or not to reconsider current intentions. With this aim, in this work, we present an abstract framework which integrates a *Possibilistic Defeasible Logic Programming* [1] approach to decision making in the inner decision processes within BDI agents.

Keywords: Agreement Technologies, Multi-criteria Decision Making, BDI, Argumentation, Possibilistic Defeasible Logic Programming

1 Introduction

The BDI model is a particular decision making model based on cognitive notions, namely: Belief, Desires and Intentions. This model is very relevant because of its similarity with human reasoning, the theoretical underpinning it has [2, 3], as well as its applicability to solve real-world problems [4, 5].

BDI architecture is inspired from Bratman's work on *practical reasoning* [2]. Practical reasoning (PR) aims at deciding what to do in a given situation and thus is directed towards action. However, besides deciding which action perform next, BDI agents also have to decide: (a) from a set of possibly conflicting desires which intention to achieve, (b) which plan execute from among the plans that satisfy the chosen intention, and (c) whether is necessary or not to reconsider current intentions. That is, BDI model also implies making background decisions.

Some of the issues mentioned above have been tackled in previous works. Casali *et al.* [6] present a general framework to define graded BDI agent architectures, where degrees in BDI models are used to set different levels of preferences or rejections on desires and preferences at intentions level to model the cost/benefit trade-off of reaching a goal. In [7], ideas from argumentation are combined with desire and planning rules, to give a formal account on how consistent sets of intentions can be obtained from a conflicting set of desires. A general framework for practical reasoning based on an abstract argumentative

machinery is presented in [8]. To the best of our knowledge, at present, there are no proposals which clearly formulate how these choices are made in BDI agent's inner decision processes. In this way, the main goal of this paper aims at incorporating in a generic way, multi-criteria decision making in BDI agent's inner decision processes. In particular, an argumentation-based approach to multi-criteria decision making is used [9]. In this respect, some proposals exist [10, 11], aiming at incorporating argumentation-based approaches within BDI agents.

The rest of the paper is organized as follows. Sect. 2 briefly introduces the BDI model to provide the background concepts underlying the proposed abstract framework (Sect. 3), which integrates multi-criteria argumentation-based decision making in the inner decision processes of the BDI architecture. Then, this framework is exemplified in the TILEWORLD domain (Sect. 4). Finally, Sect. 5 draws the conclusions and briefly describes possible future work.

2 BDI Model

Belief-Desires-Intentions models (BDI) have been inspired from the philosophical tradition on understanding *practical reasoning* and were originally proposed by Bratman *et al.* [2]. This kind of reasoning can be conceived as the process of deciding what action perform next to accomplish a certain goal. Practical reasoning involves two important processes, namely: deciding *what* states of the world to achieve and *how* to do it. The first process is known as *deliberation* and its result is a set of intentions. The second process, so-called *means-ends reasoning* involves generating actions sequences to achieve intentions.

The mental attitudes of a BDI agent on its beliefs, desires and intentions, represent its informational state, motivational state and decision state, respectively. The BDI architecture defines its cognitive notions as follows:

- **Beliefs:** Partial knowledge the agent has about the world.
- **Desires:** The states of the world that the agent would ideally like to achieve.
- **Intentions:** Desires (states of the world) that the agent has committed (dedicated resources) to achieve.

These cognitive notions are implemented as data structures in the BDI architecture, which also has an interpreter in charge of manipulating them to select the most appropriate actions to be performed by the agent. This interpreter performs the deliberation and means-ends reasoning processes aforementioned, and its simpler version is shown in Algorithm 1, as proposed in [12].

A usual problem in designing practical reasoning agents lies in getting a good balance among deliberation, means-ends reasoning and actions execution. It is clear that, in some point of time, an agent should drop some of its intentions, because they were already achieved, they are impossible to be achieved or makes no sense to do it, etc. Likewise, when opportunities arise to achieve new desires, the agent should generate intentions aiming at accomplishing them. Thus, as mentioned above it is important for an agent to *reconsider its intentions*. However, intentions reconsideration is costly in terms of time and computational resources.

Algorithm 1 Agent control loop (version 1)

```
1: while true do
2:     observe the world;
3:     update internal world model;
4:     deliberate about what intention to achieve next;
5:     use means-ends reasoning to get a plan for the intention;
6:     execute the plan;
7: end while
```

Moreover, it can happen that some of the actions from the executing plan might fail in achieving the intended results, hence *replanning* capabilities should be provided. Both replanning and intentions reconsideration (if performed) must be carried out during the execution phase of the chosen actions.

3 Integration Framework

As mentioned above, the BDI model uses the cognitive notions of beliefs, desires and intentions to decide what to do, but also, inner decisions exist related to these high-level decisions which, in our view, have not been clearly detailed in previous works. That is why, in this section we propose an abstract framework which integrates multi-criteria argumentation-based decision making to solve inner decision making in a BDI agent.

In Sect. 2 it was referred that a BDI agent comprises two fundamental processes, namely, deliberation and means-ends reasoning, which are followed by a plan execution stage. Within these processes (deliberation, means-ends reasoning and execution) the following inner decisions can be made:

- Choice among conflicting desires: *deliberation* requires to commit to an intention from among conflicting desires.
- Choice between plans: during *means-ends reasoning* it might be necessary to choose from among plans which achieve the same intention, that is, deciding which action perform to achieve a particular intention.
- Intentions reconsideration: during the *execution* process (of only one plan or a mega-plan involving all the plans the agent has committed to) decisions should be made with respect to whether reconsider or not current intentions based on the dynamics of the environment, and if so, if new intentions should be adopted or current intentions should be dropped.

All in all, our BDI architecture will incorporate an *Inner Decision Making Component (IDMC)* which will make inner decisions with respect to the different alternatives and the multiple criteria provided to the agent. In our proposal, to select the best alternative from a given set of alternatives, the agent will have the $select(\cdot,\cdot,\cdot)$ function that will return the choice made by IDMC. This function will be used (within this framework) in all the inner decision processes a BDI agent has. It will receive as input parameters: (1) a set B of candidate

alternatives, (2) the set C containing the criteria that will be used to compare alternatives among each other, and (3) the preferences $\mathcal{P}$, composed by a preference order among criteria and a preference order among the possible values an alternative can take for each particular criterion. To select an alternative, this function implements the argumentation-based decision framework proposed in [9]. Therefore, next section briefly describes this framework and a pseudo-code of the $select(\cdot,\cdot,\cdot)$ function is presented.

3.1 The Argumentation-Based Decision Framework

The argumentation-based decision framework described in this section is formally related to the *choice-based approach* (CBA) to decision making, as stated in [9]. The CBA takes as primitive object the choice behaviour of the individual, which is represented by means of a *choice structure* ($\mathcal{B}$, $\mathbf{C}(\cdot)$) consisting of two elements:

- $\mathcal{B}$ is a set of subsets of X (the set containing all the available alternatives to the decision maker). Each set $B \in \mathcal{B}$, represents a set of alternatives (or *choice experiment*) that can be conceivably posed to the decision maker.
- $\mathbf{C}(\cdot)$ is a *choice rule* which basically assigns to each set of alternatives $B \in \mathcal{B}$ a non-empty set that represents the alternatives that the decision maker *might* choose when presented the alternatives in B. ($\mathbf{C}(B) \subseteq B$ for every $B \in \mathcal{B}$). When $\mathbf{C}(B)$ contains a single element, this element represents the *individual's choice* among the alternatives in B. The set $\mathbf{C}(B)$ might, however, contain more than one element and in this case they would represent the *acceptable alternatives* in B for the decision maker.

This decision framework is conceptually composed by three components. The first component is set X. The second component, the epistemic component, represents the agent's knowledge and preferences, and the third one is the decision component. Formally, the argumentation-based decision framework is a triple $\langle X, \mathcal{K}, \Gamma \rangle$ where:

- X is the set of all the *possible alternatives* that can be presented to the decision maker.
- $\mathcal{K}$ is the *epistemic component* of the decision maker (see Definition 4.5 from [9]). Formally, $\mathcal{K}$ is a 5-tuple, $\mathcal{K} = \langle \mathcal{C}, >_{\mathcal{C}}, ACC, \Pi, \Delta \rangle$ where:
  - $\mathcal{C}$ is a set of *comparison literals* representing the preference criteria that the decision maker will use to compare the elements in X. Let $C = \{C_1, \ldots, C_n\}$ $(n > 0)$ be the set of preference criteria that will be used to compare the elements in X, each criterion C_i has associated a *comparision literal* $c_i(\cdot,\cdot)$ that states the preference between two alternatives of X, based on their attribute values. Then, $\mathcal{C}$= $\{c_1(\cdot,\cdot), \ldots, c_n(\cdot,\cdot)\}$.
  - $>_{\mathcal{C}}$ is a strict total order over the elements of $\mathcal{C}$. (Definition 4.2 from [9]).
  - ACC is a user-specified *aggregation function* that aggregate necessity degrees. ACC must satisfy specific properties (see [9]) and function $f^{+}_{\Phi}(\cdot)$ is defined from it. Here we will use:

$$ACC(\alpha_1, ..., \alpha_n) = [1 - \prod_{i=1}^{n}(1 - \alpha_i)] + k \max(\alpha_1, ..., \alpha_n) \prod_{i=1}^{n}(1 - \alpha_i) \quad (1)$$

with $k \in (0, 1)$, which has been shown in [9] to satisfy the desired properties to apply the framework.

* Π is a set of *certain clauses*.
* Δ is a set of *uncertain clauses*.

$P(\Pi, \Delta)$ is a conformant P-DeLP program (see Definition 4.3 from [9]).

– Γ is the *decision component*. It is a set with two decision rules:[1]

$$\Gamma = \left\{ \begin{array}{l} DR1 : \{W\} \stackrel{B}{\Leftarrow} \{bt(W,Y)\},\ not\{bt(Z,W)\} \\ DR2 : \{W,Y\} \stackrel{B}{\Leftarrow} \{sp(W,Y)\},\ not\{bt(Z,W)\} \end{array} \right\} \text{ with } B \subseteq X.$$

Rule DR1 states that an alternative $W \in B$ will be chosen, if W is better than another alternative Y and there is not a better alternative Z than W. Besides, rule DR2 says that two alternatives W, $Y \in B$ with the same properties will be chosen if there is not a better alternative Z than W and Y.

Let $B \in \mathcal{B}$ be a set of alternatives posed to the agent and $\langle X, \mathcal{K}, \Gamma \rangle$ be the agent's decision framework. Let $\{D_i \stackrel{B}{\Leftarrow} P_i, not\ T_i\}_{i=1...n} \subseteq \Gamma$ be the set of applicable decision rules with respect to $\mathcal{K}$. The set of *acceptable alternatives* of the agent will be $\Omega_B = \bigcup_{i=1}^{n} D_i$.

In Algorithm 2, a general algorithm which implements a choice rule $\mathbf{C}(\cdot)$ is presented. As it can be observed function μ has as input parameter a choice experiment (B). A choice experiment is a set containing at least one element, hence, this function returns *failure* if receives as argument an empty set (step 1). If the choice experiment has one element, then it is thus returned as solution since there is only one trivial choice to be made (step 2). Then, if a non-empty set was received as parameter, the resulting set *sol* is initialized (step 3) and a local copy (*ch*) of the original choice experiment is made (step 4). The computing process to determine the set of acceptable alternatives ends when *ch* becomes empty (step 6), thus exiting the main loop (step 5) returning the computed set of acceptable alternatives *sol* (step 13). While there are alternatives in *ch*, an alternative is removed from this set and is assigned to h (step 7). If there is not a better alternative than h in the choice experiment (step 9) and h is better than any other alternative in the choice experiment (step 8), then h is added to the resulting set *sol* (step 10), otherwise is discarded (step 9). Besides, if h is not better than any other alternative in the choice experiment (step 8), but there is no other alternative (let us denoted it as h') in the choice experiment better than h (step 11), then it holds that h and h' have the same properties, and they are the best, therefore h is added to the resulting set *sol* (step 12). It is worth mentioning, that in turn (when selected in step 7) h' will also be added to *sol*.

Based on the above-mentioned framework, function $select(\cdot, \cdot, \cdot)$ executes the steps shown in Algorithm 3 to choose from among the alternatives in B.

[1] Due to space constraints, the literals $better(\cdot, \cdot)$ and $same\_prop(\cdot, \cdot)$ in [9], will be referred in this paper as $bt(\cdot, \cdot)$ and $sp(\cdot, \cdot)$, respectively.

Algorithm 2 Compute Acceptable Alternatives

function μ(choice-experiment) **returns** non-empty-set-of-alternatives, or failure
1: **if** EMPTY?(choice-experiment) **then return** failure
2: **if** SINGLETON?(choice-experiment) **then return** choice-experiment
3: $sol \leftarrow \emptyset$
4: $ch \leftarrow$ choice-experiment
5: **loop do**
6: **if** EMPTY?(ch) **then exit**
7: $h \leftarrow$ REMOVE-ELEMENT(ch)
8: **if** IS-h-BETTER-THAN-ANY-OTHER?(choice-experiment) **then**
9: **if** ANY-BETTER-THAN-h?(choice-experiment) **then** discard h
10: **else** ADD-ELEMENT(sol,h)
else
11: **if** ANY-BETTER-THAN-h?(choice-experiment) **then** discard h
12: **else** ADD-ELEMENT(sol,h)
13: **return** sol

Algorithm 3 Computation for alternatives selection

function *select*(alternatives B, criteria C, preferences $\mathcal{P}$) **returns** non-empty-set-of-alternatives, or failure

1: Define the comparison literal for each $C_i \in C$
2: Define $>_C$ according to the preferences of each criterion in $\mathcal{P}$
3: Build a conformant program $P(\Pi, \Delta)$ (as defined in [9])
4: **return** Evaluation of function $\mu(B)$

4 Example: The Tileworld

The TILEWORLD experimental domain [13] is a grid environment containing agents, tiles, holes and obstacles. The agent's objective consists of scoring as many points as possible by pushing the tiles into the holes to fill them in. The agent is able to move up, down, left, or right, one cell at a time, having as only restriction that obstacles must be avoided. This environment is dynamic, so that holes and tiles may randomly appear and disappear in accordance to a series of world parameters, which can be varied by the experimenter.

A BDI agent for the TILEWORLD can be implemented as follows: the agent's beliefs consist of its perceptions about the objects locations, as well as the score and time-out time for all the holes. Desires are the holes to be filled in, and the current intention (IH) aims at filling a particular hole right now. The means-end reasoner basically is a special-purpose route planner, which guides the agent to a particular tile that must be pushed into the hole to be filled in. Figure 1 shows a hypothetical scene in which the framework proposed in Sect. 3 will be used.

The agent gets its perception and updates its beliefs, in order to deliberate about what intention to achieve next. During deliberation it gets its reachable holes (options), *i.e.,* those which are not surrounded by obstacles and their time-out times are higher or equal to the distances from the agent to the holes. Then,

filtering stage takes place where one of the reachable holes is selected and becomes IH. In this case, all options are conflicting each other, since it is not possible to fill in more than one hole at a time. Hence, all reachable holes will serve as input to $selec(\cdot,\cdot,\cdot)$ function. In this way, $B = \{h_3, h_4, h_5\}$, $C = \{C_1 = \text{score}, C_2 = \text{timeout}, C_3 = \text{distAgent}, C_4 = \text{tileAvail (distance to the nearest tile)}\}$ and $>_C = \{(distAgent, timeout), (distAgent, tileAvail), (timeout, tileAvail), (score, distAgent), (score, timeout), (score, tileAvail)\}$.

Figure 2 presents preference for each criterion. Table 1 shows the alternatives and their respective values for each criterion. Likewise, following the approach from [9], a conformant P-DeLP program would be:

$$\Delta = \left\{ \begin{array}{ll} (score(h_4,h_3),0.92) & (bt(W,Y) \leftarrow score(W,Y),0.99) \\ (score(h_4,h_5),0.83) & (\sim bt(W,Y) \leftarrow score(Y,W),0.99) \\ (score(h_5,h_3),0.83) & (bt(W,Y) \leftarrow distAgent(W,Y),0.74) \\ (distAgent(h_3,h_5),0.62) & (\sim bt(W,Y) \leftarrow distAgent(Y,W),0.74) \\ (distAgent(h_3,h_4),0.67) & (bt(W,Y) \leftarrow timeout(W,Y),0.49) \\ (distAgent(h_5,h_4),0.54) & (\sim bt(W,Y) \leftarrow timeout(Y,W),0.49) \\ (timeout(h_5,h_3),0.37) & (bt(W,Y) \leftarrow tileAvail(W,Y),0.24) \\ (timeout(h_5,h_4),0.38) & (\sim bt(W,Y) \leftarrow tileAvail(Y,W),0.24) \\ (timeout(h_3,h_4),0.26) & \\ (tileAvail(h_3,h_5),0.08) & \\ (tileAvail(h_4,h_5),0.08) & \end{array} \right\}$$

$$\Pi = \left\{ (\sim bt(W,Y) \leftarrow sp(W,Y),1) \qquad (\sim bt(W,Y) \leftarrow sp(Y,W),1) \right\}$$

In the particular program presented above, the necessity degrees of the clauses belonging to (Π, Δ) were calculated as follows:

1. Normalize the alternatives' attribute values to interval $[0, 1]$ for all of the preference criteria (see Table 1).
2. Compare the alternatives among each other with respect to the normalized preference criteria (see first column of Table 2). The necessity degree of the clause is calculated as the absolute value of the remainder of their normalized attribute values.
3. Divide the necessity degrees obtained in previous step by the number of preference criteria provided to the decision maker, *i.e.,* by 4 in this case (see second column of Table 2).
4. Map the necessity degrees obtained in previous step to the subinterval assigned to the comparison literal in the clause (see third column of Table 2).
5. For each clause (φ, α) such that φ is a rule of the kind $bt(W,Y) \leftarrow c_i(W,Y)$ or $\sim bt(W,Y) \leftarrow c_i(Y,W)$, set α to the upper bound value of the subinterval assigned to $c_i(\cdot,\cdot)$.

Alternatives	C_1	C_2	C_3	C_4	$C_{1[0,1]}$	$C_{2[0,1]}$	$C_{3[0,1]}$	$C_{4[0,1]}$
$h3$	3	8	2	2	0.33	0.53	0.33	0.67
$h4$	9	7	6	2	1	0.47	1	0.67
$h5$	6	15	5	3	0.67	1	0.83	1

Table 1. Alternatives values for each criterion

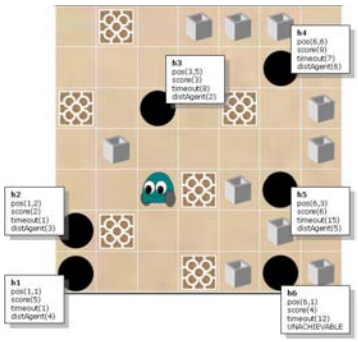


Fig. 1. Tileworld scene

Criteria	Comparison literal	Subinterval
C_1	$score(\cdot,\cdot)$	$[0.75, 1)$
C_2	$timeout(\cdot,\cdot)$	$[0.25, 0.5)$
C_3	$distAgent(\cdot,\cdot)$	$[0.50, 0.75)$
C_4	$tileAvail(\cdot,\cdot)$	$[0, 0.25)$

Fig. 2. Preferences per criterion

$(score(h_4,h_3),0.67)$	$(score(h_4,h_3),0.17)$	$(score(h_4,h_3),0.92)$
$(score(h_4,h_5),0.33)$	$(score(h_4,h_5),0.08)$	$(score(h_4,h_5),0.83)$
$(score(h_5,h_3),0.34)$	$(score(h_5,h_3),0.08)$	$(score(h_5,h_3),0.83)$
$(distAgent(h_3,h_5),0.5)$	$(distAgent(h_3,h_5),0.12)$	$(distAgent(h_3,h_5),0.62)$
$(distAgent(h_3,h_4),0.67)$	$(distAgent(h_3,h_4),0.17)$	$(distAgent(h_3,h_4),0.67)$
$(distAgent(h_5,h_4),0.17)$	$(distAgent(h_5,h_4),0.04)$	$(distAgent(h_5,h_4),0.54)$
$(timeout(h_5,h_3),0.47)$	$(timeout(h_5,h_3),0.12)$	$(timeout(h_5,h_3),0.37)$
$(timeout(h_5,h_4),0.53)$	$(timeout(h_5,h_4),0.13)$	$(timeout(h_5,h_4),0.38)$
$(timeout(h_3,h_4),0.06)$	$(timeout(h_3,h_4),0.01)$	$(timeout(h_3,h_4),0.26)$
$(tileAvail(h_3,h_5),0.33)$	$(tileAvail(h_3,h_5),0.08)$	$(tileAvail(h_3,h_5),0.08)$
$(tileAvail(h_4,h_5),0.33)$	$(tileAvail(h_4,h_5),0.08)$	$(tileAvail(h_4,h_5),0.08)$

Table 2. Alternatives comparison

The following arguments are built from the above P-DeLP conformant program:[2]

$\mathcal{A}_1 = \{(bt(h_4,h_3) \leftarrow score(h_4,h_3),0.99),\ (score(h_4,h_3),0.92)\}$
$\mathcal{A}_2 = \{(\sim bt(h_3,h_4) \leftarrow score(h_4,h_3),0.99),\ (score(h_4,h_3),0.92)\}$
$\mathcal{A}_3 = \{(bt(h_4,h_5) \leftarrow score(h_4,h_5),0.99),\ (score(h_4,h_5),0.83)\}$
$\mathcal{A}_4 = \{(\sim bt(h_5,h_4) \leftarrow score(h_4,h_5),0.99),\ (score(h_4,h_5),0.83)\}$
$\mathcal{A}_5 = \{(bt(h_5,h_3) \leftarrow score(h_5,h_3),0.99),\ (score(h_5,h_3),0.83)\}$
$\mathcal{A}_6 = \{(\sim bt(h_3,h_5) \leftarrow score(h_5,h_3),0.99),\ (score(h_5,h_3),0.83)\}$
$\mathcal{A}_7 = \{(bt(h_3,h_5) \leftarrow distAgent(h_3,h_5),0.74),\ (distAgent(h_3,h_5)),0.62)\}$
$\mathcal{A}_8 = \{(\sim bt(h_5,h_3) \leftarrow distAgent(h_3,h_5),0.74),\ (distAgent(h_3,h_5)),0.62)\}$
$\mathcal{A}_9 = \{(bt(h_3,h_4) \leftarrow distAgent(h_3,h_4),0.74),\ (distAgent(h_3,h_4),0.67)\}$
$\mathcal{A}_{10} = \{(\sim bt(h_4,h_3) \leftarrow distAgent(h_3,h_4),0.74),\ (distAgent(h_3,h_4),0.67)\}$
$\mathcal{A}_{11} = \{(bt(h_5,h_4) \leftarrow distAgent(h_5,h_4),0.74),\ (distAgent(h_5,h_4),0.54)\}$
$\mathcal{A}_{12} = \{(\sim bt(h_4,h_5) \leftarrow distAgent(h_5,h_4),0.74),\ (distAgent(h_5,h_4),0.54)\}$
$\mathcal{A}_{13} = \{(bt(h_5,h_3) \leftarrow timeout(h_5,h_3),0.49),\ (timeout(h_5,h_3),0.37)\}$
$\mathcal{A}_{14} = \{(\sim bt(h_3,h_5) \leftarrow timeout(h_5,h_3),0.49),\ (timeout(h_5,h_3),0.37)\}$
$\mathcal{A}_{15} = \{(bt(h_5,h_4) \leftarrow timeout(h_5,h_4),0.49),\ (timeout(h_5,h_4),0.38)\}$
$\mathcal{A}_{16} = \{(\sim bt(h_4,h_5) \leftarrow timeout(h_5,h_4),0.49),\ (timeout(h_5,h_4),0.38)\}$

[2] To simplify notation, given an argument $\langle \mathcal{A}, h, \alpha \rangle$, only the set $\mathcal{A}$ of uncertain clauses will be given since the conclusion h and its necessity degree α can be obtained from it.

$\mathcal{A}_{17} = \{(bt(h_3, h_4) \leftarrow timeout(h_3, h_4), 0.49),\ (timeout(h_3, h_4), 0.26)\}$
$\mathcal{A}_{18} = \{(\sim bt(h_4, h_3) \leftarrow timeout(h_3, h_4), 0.49),\ (timeout(h_3, h_4), 0.26)\}$
$\mathcal{A}_{19} = \{(bt(h_3, h_5) \leftarrow tileAvail(h_3, h_5), 0.24),\ (tileAvail(h_3, h_5), 0.08)\}$
$\mathcal{A}_{20} = \{(\sim bt(h_5, h_3) \leftarrow tileAvail(h_3, h_5), 0.24),\ (tileAvail(h_3, h_5), 0.08)\}$
$\mathcal{A}_{21} = \{(bt(h_4, h_5) \leftarrow tileAvail(h_4, h_5), 0.24),\ (tileAvail(h_4, h_5), 0.08)\}$
$\mathcal{A}_{22} = \{(\sim bt(h_5, h_4) \leftarrow tileAvail(h_4, h_5), 0.24),\ (tileAvail(h_4, h_5), 0.08)\}$

To calculate the accrued structures for these arguments, it will be used the ACC function defined below, with $K = 0.1$:[3]

$$ACC(\alpha_1, \ldots, \alpha_n) = [1 - \textstyle\prod_{i=1}^{n}(1 - \alpha_i)] + K \max(\alpha_1, \ldots, \alpha_n) \textstyle\prod_{i=1}^{n}(1 - \alpha_i)$$

As it can be observed, twelve a-structures can be built to support the reasons by which an alternative should be deemed better than another one.

$[\Phi_1, bt(h_3, h_5), 0.67]$, $[\mathbf{\Phi'_1}, \mathbf{\sim bt(h_3, h_5), 0.90}], \Phi_1 = \mathcal{A}_7 \cup \mathcal{A}_{19},\ \Phi'_1 = \mathcal{A}_6 \cup \mathcal{A}_{14};$
$[\Phi_2, \sim bt(h_5, h_3), 0.67], [\mathbf{\Phi'_2}, \mathbf{bt(h_5, h_3), 0.90}]$, $\Phi_2 = \mathcal{A}_8 \cup \mathcal{A}_{20},\ \Phi'_2 = \mathcal{A}_5 \cup \mathcal{A}_{13};$
$[\Phi_3, bt(h_3, h_4), 0.78]$, $[\mathbf{\Phi'_3}, \mathbf{\sim bt(h_3, h_4), 0.93}], \Phi_3 = \mathcal{A}_9 \cup \mathcal{A}_{17},\ \Phi'_3 = \mathcal{A}_2;$
$[\Phi_4, \sim bt(h_4, h_3), 0.78], [\mathbf{\Phi'_4}, \mathbf{bt(h_4, h_3), 0.93}]$, $\Phi_4 = \mathcal{A}_{10} \cup \mathcal{A}_{18}, \Phi'_4 = \mathcal{A}_1;$
$[\Phi_5, bt(h_5, h_4), 0.73]$, $[\mathbf{\Phi'_5}, \mathbf{\sim bt(h_5, h_4), 0.85}], \Phi_5 = \mathcal{A}_{11} \cup \mathcal{A}_{15}, \Phi'_5 = \mathcal{A}_4 \cup \mathcal{A}_{22};$
$[\Phi_6, \sim bt(h_4, h_5), 0.73], [\mathbf{\Phi'_6}, \mathbf{bt(h_4, h_5), 0.85}]$, $\Phi_6 = \mathcal{A}_{12} \cup \mathcal{A}_{16}, \Phi'_6 = \mathcal{A}_3 \cup \mathcal{A}_{21};$

Those a-structures warranted from the dialectical process (shown in bold), will be used by Algorithm 2 to compute the set of acceptable alternatives. In this particular case, only decision rule DR1 can be applied. For alternative h_4, precondition of DR1 can be warranted and like there is no warranted a-structure supporting a conclusion of the kind $bt(Z, h_4)$ to warrant DR1's restriction, h_4 becomes the acceptable alternative. Finally, hole h_4 becomes IH.

Once a hole has been selected to fill in, plans to achieve this intention are selected. The criteria set provided for plan selection could be $C = \{C_1 = length$, $C_2 = \text{cost}$, $C_3 = \text{timeoutTile}\}$. Criterion C_1 is the number of action within the plan. C_2 represents the plan cost which is calculated as the sum of its actions costs, which depend on the agent's orientation. Finally, C_3 is the time-out time of the tile selected in the plan to fill in the hole.

On the other hand, the fact that holes appear and disappear causes the agent to change its intentions. For example, when the set of holes dot not change while the agent is executing a plan, then there is no need to deliberate; but if the set of holes do change, this might mean that IH has disappeared or that a closer hole has appeared; thus, intentions reconsideration is necessary. To achieve this behaviour, it is important to consider appropriate criteria to determine whether these changes have occurred or not. Means-ends reasoning and intention reconsideration also use the argumentation-based decision framework (as in the filtering stage), in order to choose a plan to execute or to reconsider intentions, while the plan is under execution. Due to space constraints, how this framework is applied in these stages, will not be developed in this paper.

[3] This function is a variant of the One-Complement accrual function used in [14] where K aims at weighting the importance given to the highest priority preference criterion.

5 Conclusions

In this work, we presented an abstract framework that integrates argumentation-based decision making from a multi-criteria approach, within the inner decision processes of a BDI agent. In this way, the contribution of this work is twofold. On one hand, it was specified how to perform concrete implementations of inner decision making processes within a BDI agent. On the other hand, different criteria and preferences were aggregated to get a solution to a multi-criteria decision problem as an instantiation of argumentation-based decision making.

In order to get a better understanding and provide feedback to the abstraction process carried out to propose this present framework, as future work, following the idea proposed in [15], new instantiations of the framework will be done with other methods belonging to the research field of Agreement Technologies.

References

1. Alsinet, T., Chesñevar, C.I., Godo, L., Simari, G.: A logic programming framework for possibilistic argumentation: formalization and logical properties. Fuzzy Sets and Systems **159**(10) (2008) 1208–1228
2. Bratman, M., Israel, D., Pollack, M.: Plans and resource bounded reasoning. Computational Intelligence **4**(4) (1988) 349–355
3. Dennett, D.C.: Intentional systems. Journal of Philosophy **68** (1971) 87–106
4. Evertsz, R., Fletcher, M., Jones, R., Jarvis, J., Brusey, J., Dance, S.: Implementing Industrial Multi-agent Systems Using JACK. In: Programming Multi-Agent Systems. Springer (2004)
5. Benfield, S.S., Hendrickson, J., Galanti, D.: Making a strong business case for multiagent technology. In: 5th AAMAS. (2006)
6. Casali, A., Godo, L., Sierra, C.: A graded BDI agent model to represent and reason about preferences. Artifical Intelligence **175**(7-8) (2011) 1468–1478
7. Amgoud, L.: A formal framework for handling conflicting desires. In Nielsen, T.D., Zhang, N.L., eds.: ECSQARU. Volume 2711 of Lecture Notes in Computer Science., Springer (2003) 552–563
8. Amgoud, L., Prade, H.: Formalizing practical reasoning under uncertainty: An argumentation-based approach. In: IAT, IEEE Computer Society (2007) 189–195
9. Ferretti, E., Errecalde, M., García, A., Simari, G.: A possibilistic defeasible logic programming approach to argumentation-based decision making. Manuscript ID: TETA-2012-0093.R1. Under Review process in *JETAI*. `https://sites.google.com/site/edgardoferretti/TETA-2012-0093.R1.pdf?attredirects=0&d=1`.
10. Rotstein, N.D., García, A.J., Simari, G.R.: Reasoning from desires to intentions: A dialectical framework. In: AAAI, AAAI Press (2007) 136–141
11. Schlesinger, F., Ferretti, E., Errecalde, M., Aguirre, G.: An argumentation-based BDI personal assistant. In: IEA/AIE. Volume 6069 of LNAI., Springer (2010)
12. Wooldridge, M.: Reasoning about Rational Agents. The MIT Press (2000)
13. Pollack, M.E., Ringuette, M.: Introducing the tileworld: Experimentally evaluating agent architectures. In: 8th AAAI. (1990) 183–189
14. Gómez, M., Chesñevar, C., Simari, G.: Modelling argument accrual in possibilistic defeasible logic programming. In: ECSQARU. LNCS, Springer (2009) 131–143
15. Sosa-Toranzo, C., Schlesinger, F., Ferretti, E., Errecalde, M.: Integrating a voting protocol within an argumentation-based BDI system. In: XVI CACIC. (2010)

Una Extensión de Agentes en JASON para Razonar con Incertidumbre: G-JASON

Adrián Biga[1] and Ana Casali[1,2]

[1] Facultad de Cs. Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario (UNR)
Av. Pellegrini 250 - S2000BTP, Rosario, Argentina
[2] Centro Internacional Franco Argentino de
Ciencias de la Información y de Sistemas (CIFASIS)
aebiga@gmail.com, acasali@fceia.unr.edu.ar

Abstract. Una de las mejores implementaciones de agentes de la familia BDI (B: Belief, D: Desire, I: Intention) es mediante los llamados Sistemas de Razonamiento Procedural (PRS). En este trabajo se plantea una extensión de los PRS para permitir generar agentes más flexibles, que puedan representar la incertidumbre del entorno y distintos grados de relevancia en los planes del agente. La extensión propuesta se implementó en la plataforma JASON, que permite la implementación de agentes PRS en JAVA otorgándoles alta portabilidad.

Keywords: Agentes de Razonamiento Procedural, JASON, modelo BDI, incertidumbre

1 Introducción

En los últimos años ha crecido el interés en modelar sistemas complejos como sistemas multiagentes [10]. Dentro de las arquitecturas más notorias para dar soporte a los agentes que componen estos sistemas, se encuentra la arquitectura BDI (B: Belief, D: Desiere, I: Intention)([9],[7]). Esta arquitectura de agentes ha sido una de las más estudiada y utilizada en distintas aplicaciones reales de importancia [4].

Los Sistemas de Razonamiento Procedural (PRS) [6], son las implementaciones más conocidas de una arquitectura basada en el paradigma BDI. Desde la primera implementación de un sistema PRS [5], se han desarrollado distintas versiones, respecto a implementaciones en Java destacamos dos versiones actualmente utilizadas: JACK [3] y JASON [1].

Una de las limitaciones tanto el modelo BDI definido por Rao y Georgeff [9] como los sistemas basados en PRS [5], es que no contemplan una forma explícita de representar la incertidumbre del entorno, estos se han planteado considerando una lógica bi-valuada para dar soporte a toda la información que utiliza un agente para la toma de decisiones (representada en sus estados mentales: B, D e I).

[3] http://aosgrp.com/products/jack/

En el modelo BDI graduado (g-BDI) presentado en Casali et al.[3, 2], se plantea un modelo general de agente BDI especificando una arquitectura que pueda tratar con la incertidumbre del entorno y actitudes mentales graduadas. De esta forma los grados de las creencias van a representar en que medida el agente cree que una fórmula es cierta, los grados de los deseos permiten representar el nivel de preferencia y el grado de las intenciones darán una medida costo-beneficio que le representa al agente alcanzar cada objetivo. Para especificar la arquitectura de un agente BDI graduado, se utiliza la noción de sistema multicontextos. De esta forma, un agente g-BDI es definido como un grupo de contextos interconectados para representar sus creencias, deseos e intenciones. Cada contexto tiene su lógica asociada, para representar y razonar con grados en las creencias, deseos e intenciones, se ha elegido utilizar lógicas modales multivaluadas. Las características generales de los distintos componentes de un agente BDI graduado se pueden ver en [3]. Se ha observado que el modelo g-BDI brinda un framework que permite modelizar agentes más flexibles y que pueden representar apropiadamente la incertidumbre del entorno y preferencias del agente [4], pero aún no se ha desarrollado una plataforma que permita una implementación genérica de los agentes especificados bajo este modelo.

Con inspiración en este modelo de agentes g-BDI, en este trabajo se presenta una extensión de los sistemas PRS para representar estructuras de datos graduadas en sus creencias (Beliefs) y en los planes del agente.

2 Arquitecturas de Agentes y Sistemas PRS

Existen diferentes propuestas para la clasificación de arquitecturas de agentes. Tomando como referencia la clasificación realizada por Wooldrige [10], se puede considerar las siguientes clases concretas de arquitecturas: deliberativas, reactivas, híbridas y arquitecturas de razonamiento práctico. Dentro de este último grupo se encuentran los agentes BDI.

La arquitectura BDI está caracterizada porque los agentes están dotados de estructura de datos que representan explícitamente los siguientes estados mentales:

- Creencias (Beliefs): representan el conocimiento que el agente tiene sobre sí mismo y sobre el entorno.
- Deseos (Desires): son los objetivos que el agente desea cumplir.
- Intenciones (Intentions): se puede considerar como un subconjunto de deseos consistentes entre sí que el agente decide alcanzar. Las intenciones derivan en las acciones que ejecutará el agente en cada momento.

La arquitectura PRS desarrollada por Georgeff y Lansky [6] fue quizás la primera arquitectura basada en el paradigma BDI. Ha sido utilizada en varias aplicaciones ([7],[8]). Un agente con arquitectura PRS trata de alcanzar cualquier meta que se proponga basándose en sus creencias sobre el mundo (entorno). También puede simultáneamente reaccionar ante la ocurrencia de algún nuevo evento. De esta forma, PRS provee un marco en el cual los comportamientos de tipo *dirigido a la meta* y *dirigido por los eventos* pueden ser fácilmente integrados. Los sistemas

PRS consisten en un conjunto de herramientas y métodos, para la representación y ejecución de planes. Estos planes o procedimientos son secuencias condicionales de acciones las cuales pueden ejecutarse para alcanzar ciertos objetivos, o reaccionar en situaciones particulares.

3 La Arquitectura de JASON

El lenguaje interpretado por JASON es una extensión del AgentSpeak(L) [1] , un lenguaje abstracto que tiene una concisa notación y es una eficiente extensión de la programación lógica para sistemas BDI. JASON a comparación de otros sistemas de agentes BDI posee la ventaja de ser multiplataforma al estar desarrollado en el lenguaje JAVA.

Un agente en **AgentSpeak (L)** es creado especificando un conjunto de creencias (beliefs) y un conjunto de planes (plans). Otros elementos relevantes son los objetivos (goals) del agente y los eventos disparadores (trigger events) que sirven para representar la parte reactiva de un agente.

- **Creencias:** representan las creencias del agente respecto a su entorno.
- **Objetivos:** representan los objetivos del agente, AgentSpeak (L) distingue sólo dos tipos de objetivos (goals): *achivement goals* y *test goals*. El *achivement goal* denota que el agente quiere alcanzar un estado en el mundo donde el predicado asociado sea verdadero. Un *test goal* devuelve una unificación asociada con un predicado en el conjunto de creencias del agente, si no hay asociación, simplemente falla.
- **Evento disparador** (*trigger event*) es un evento que puede iniciar la ejecución de un plan. Un evento puede ser interno, cuando un sub-objetivo tiene que ser logrado, o externo, cuando es generado por actualizaciones de creencias debido a una percepción del ambiente.
- **Planes:** son acciones básicas que un agente puede realizar sobre su ambiente. Un plan está formado por un evento disparador (denotando el propósito del plan), seguido por una conjunción de literales de creencia representando un contexto. Para que el plan sea aplicable, el contexto debe ser una consecuencia lógica de las actuales creencias del agente. El resto del plan es una secuencia de acciones básicas o subobjetivos que el agente tiene que lograr (o testear) cuando el plan es elegido para su ejecución.

La sintaxis de AgentSpeak(L) está definida por la gramática que se muestra en la Figura 1. Un agente *ag* es especificado como un conjunto de creencias *bs* (la base de creencia inicial) y un conjunto de planes *ps* (la librería de planes del agente). Un plan esta definido por $p ::= te : ct < -h\ p$, donde *te* es un evento disparador, *ct* es el contexto del plan y *h* es una secuencia de acciones, objetivos, o actualizaciones de creencias; *te:ct* es la cabeza del plan y *h* es el cuerpo.

La Figura 2 describe cómo trabaja un intérprete de AgentSpeak (L). En cada ciclo de interpretación de un programa de agente, se actualiza la creencia del agente y con ello la lista de eventos, que pudo ser generado por una percepción del ambiente (evento externo) o por la ejecución de intenciones(acciones), lo cual

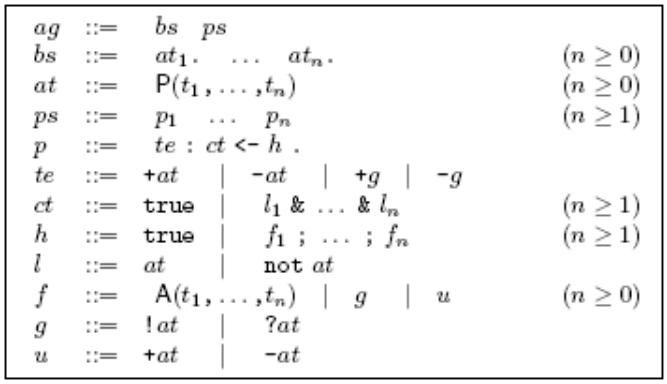


Fig. 1. Sintaxis de AgentSpeak
Fuente: [1]

produce un evento interno (ver (1) en la Figura 2). Después de que la *función Selectora de Eventos* (SE) haya seleccionado el evento (2), el agente tiene que unificar este evento con eventos disparadores en la cabeza de los planes (3). Esto genera un conjunto de planes relevantes con sus respectivos contextos, se eliminan aquellos cuyos contextos no se satisfagan con la base de creencias del agente, para formar un conjunto de planes aplicables (4). Este conjunto de planes se denominan "'opciones"'. Estas opciones se generaron a partir de un evento externo o interno. Entonces, la *función Selectora de Opciones SO* elige uno de estos planes (5), luego pone este plan con alguna intención existente (si el evento fue interno) o crea una nueva intención (si el evento fue externo). Todo lo que queda es seleccionar una intención (6) para ser ejecutada en el ciclo (7). La función Selectora de intenciones (SI) se encarga de esto. La Gramática de Jason presenta mejoras respecto a la de AgentSpeak (L) y se puede ver su definición completa en [1]. Siempre se deberá tener en cuenta que el estado de un agente, a lo largo de su ciclo de vida, está representado por un conjunto de creencias y un conjunto de planes y su comportamiento estará reflejado en el comportamiento de las funciones selectoras.

Una de las diferencias más importante que distingue a JASON de la sintaxis de AgentSpeak(L) (ver Figura 1) es la inclusión de las anotaciones tanto en las creencias como en los planes. Las anotaciones en las creencias y en los planes pueden generarse por $atomic\_formula$ mediante las siguientes reglas BNF:

$$beliefs- > (literal \text{“.”})^*; (3.1)$$
$$literal- > [\sim] atomic\_formula \ (3.2)$$
$$plan- > [@atomic\_formula] triggering\_event [: context] [\text{“} < - \text{”} body] \text{“.”} \ (3.3)$$
$$atomic\_formula- > (< ATOM > | < VAR >) [\text{“(”} list\_of\_terms \text{“)”}] [\text{“[”} list\_of\_terms \text{“]”}]; (3.4)$$

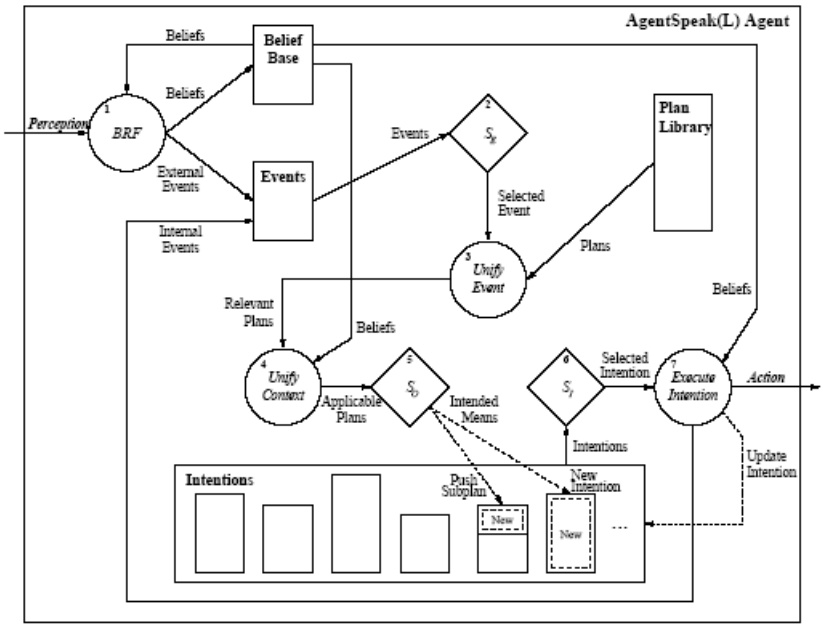


Fig. 2. Diagrama del intérprete de un agente AgentSpeak(L)
Fuente: [1]

Las fórmulas atómicas pueden contener anotaciones, esta es una lista de términos encerradas en corchetes inmediatamente siguiendo una fórmula (ver regla 3.4). Los planes tienen etiquetas representados por [@$atomic\_formula$] (ver regla 3.3).

4 Extensión G-JASON

Con la inspiración del modelo g-BDI, se presenta una extensión sintáctica y semántica de JASON que permita representar grados en las creencias y en los planes, ya que estos son los elementos fundamentales en su arquitectura:

1. *Grados en las creencias:* se agregan valores numéricos en el interválo [0,1] para representar el grado de certeza de que un hecho sea cierto, para ello se utilizan las anotaciones. Luego, una creencia queda definida por:
   $X[degOfCert(valor_x)]$ donde X representará una fórmula atómica (representando un hecho) y $valor_x$ su grado de certeza.
   Por ejemplo, la creencia de que hay un viento leve con una certeza de 0.7 se representará como: viento_leve [degOfCert(0.7)].
   Este grado impactará en el orden de selección de los eventos debido a que la función selectora SE tomará el de mayor grado primero.
2. *Grados en los Planes:* el plan está formalizado por:
   $@label[planRelevance(gradoLabel)]te[degOfCert(gradoTE)] :$
   $ct[degOfCert(gradoCT)] <- body;$

Se permite asociar a cada plan tres valores: el grado en la anotación del label (gradoLabel), en el evento disparador (gradoTE) y en el contexto (gradoCT):

- *gradoLabel* representa la medida de éxito del plan (planRelevance), este valor influye en la selección de intenciones y en la selección de opciones. De modo de que si dos planes han pasado los filtros debido a las unificaciones se ejecutará el que tenga el *gradoLabel* más alto.
- *gradoTE* es un grado aplicado al evento disparador e influye en la selección de eventos y en la unificación de eventos.
- *gradoCT* es un grado en el contexto del plan e influirá en la unificación del contexto.

Tanto el gradoTE aplicado al evento disparador, como el gradoCT aplicado al contexto, representan precondiciones adicionales: en que grado estas creencias deben ser ciertas en la base de creencias actual para que el plan sea aplicable.

Al incluir los grados mencionados, para que los mismos tengan el valor semántico deseado, se tienen que modificar las funciones selectoras descriptas y los procesos de unificación para los eventos y los contextos de los planes. A continuación se verá como se extendió la sintaxis de JASON para obtener las nuevas funcionalidades.

4.1 La Gramática de G-JASON

Las palabras "planRelevance" (relevancia de un plan) y "degOfCert" (grado de certeza) serán palabras reservadas y los grados serán representados por números flotante en el interválo [0,1]. Se modifica la sintaxis original de JASON para las creencias, planes, eventos disparadores (te) y contextos (ct):

1. Creencias: se extienden para permitir que tengan un grado de certeza:
   $beliefs ->$ (**literalC** ".") $*$ (*en la sintaxis original* *regla* 3.1)
   $literalC ->$ [" $\sim$ "]$atomic\_formula$ **anotC**
   $anotC ->$ "[$degOfCert$(" $< valor >$ ")]"
   $valor -> dig1, dig2$
   $dig1 -> 0|1$
   $dig2 -> 0|1|...|9$
2. Evento disparador (TE) del plan: se permite que sea graduado, para ello se especializa la anotación para representar el grado de certeza:
   $triggering\_event ->$ (" + " | " − ") ["!" | "?"] **literalC**
3. Contexto: se realiza el cambio de manera más directa reutilizando **anotC**
   $context ->$ log_expr **anotC** | "*true*"
4. Planes: se agrega el grado (planRelevance) dentro del plan
   $plan ->$ "@" $atomic\_formula$ **anotG** $triggering\_event$ [: $context$]
   [" $< -$"$body$] "."
   $anotG ->$ "[$planRelevance$(" $< valor >$ ")]"

4.2 Implementación de las extensiones: G-JASON

Se realizaron las modificaciones en el código de JASON tanto sobre los beliefs como sobre los planes. Las modificaciones realizadas se han implementado en las funciones selectoras y en los procesos de unificación. El código de G-JASON está disponible en `https://github.com/secharte/ab/`. Se presentan los cambios propuestos según el orden en el ciclo de interpretación de un programa de agente, siguiendo el flujo de ejecución en la máquina de estados (Figura 2).

Selección de eventos: durante la ejecución del proceso de inicialización de la máquina de estados se cargan el conjunto de eventos y el conjunto de todos los planes. El conjunto de eventos se ha generado a partir de las creencias definidas para el agente y la librería de planes se ha generado con todos los planes que contiene el agente.

1. *Opción utilizando grados en las creencias:* las creencias generan los eventos que luego unificarán con los eventos disparadores de los planes. Se decidió que la "selectora de eventos" seleccione el evento con mayor grado de certeza (degOfCert), modelando atender primero los hechos más confiables. El método recibe como parámetro una cola de eventos pero retorna el evento que posee el mayor grado de certeza (defOfCert).
   Si se generaron los eventos:
   $E1 = [+sol[degOfCert(0.8)]]$ y $E2 = [+viento[degOfCert(0.7)]]$
   La selectora retorna el evento $E1$ debido a que es el evento con mayor grado de certeza.
2. *Activando prioridades:* para manejar la reactividad del agente de una forma más directa, se agrega la posibilidad de activar un archivo "'*Prioridades*"' donde se ordenan los eventos según su prioridad. Luego, cuando esta prioridad es mayor, los eventos serán más reactivos dado que la selectora de eventos los considerará primero. Para el tratamiento de este archivo se ha modificado el proceso de selección de eventos nuevamente y se determinará en estas propiedades el orden en el cual se seleccionarán los eventos, quedando sin efecto en este caso la selección de eventos observada anteriormente.
   Por ejemplo, en una situación donde se requiere establecer el orden de importancia de tres hechos gas, luz, vibración, donde la importancia de los hechos a tratar sigue ese orden, se establece este orden de precedencia en el archivo. *Prioridades*.

Proceso de unificación de eventos: se procede a unificar el evento seleccionado por la selectora de eventos contra los eventos disparadores contenidos en los planes de la librería de planes. Si se tiene un *Plan P*:

$@PlanP + sensor\_termico(TEMP)[degOfCert(0.7)] : (TEMP < 300) < -encender\_alarma; llamar\_supervisor$,

y un evento : $E = [+sensor\_termico(100)[degOfCert(0.8)]]$:

El evento E y el plan P unifican según la unificación original de JASON: sensor_termico(100) del evento unifica con sensor_termico(TEMP) del plan y quedará TEMP=100.

Se extendió esta unificación y se seleccionarán los planes donde además, el grado del evento unificado sea mayor al grado del evento disparador del plan considerado.

En el ejemplo, el grado del evento E es 0.8 y el grado del evento disparador del $PlanP$ es 0.7. Luego este plan será considerado relevante.

Proceso de unificación de los contextos el proceso de unificación de eventos arroja un conjunto de planes relevantes los cuales serán evaluados en el proceso de unificación de contexto. Originalmente en JASON se evalúa que el contexto sea verdadero. Por ejemplo, si se tiene que el *planP* (citado anteriormente) es relevante y una creencia en la base: $B = [+sensor\_termico(100)[degOfCert(0.8)]]$:

Como observamos en la unificación anterior TEMP tomara el valor 100, con lo cual la validación del contexto (TEMP <300) será verdadera. Se extendió el proceso de unificación de contexto y se seleccionará el plan relevante donde el grado del belief en la base sea mayor al grado del contexto en el plan. En el ejemplo, el grado del belief B es 0.8 y el grado del contexto del plan relevante P es 0.6 por lo tanto el $PlanP$ se considerará *aplicable*.

Selección de opciones: la unificación de contextos arrojará un conjunto de planes aplicables, la selectora de opciones elige uno de estos planes aplicables y luego vincula este plan con alguna intención existente (en caso de el evento fue interno, generado a partir de una creencia) o se creará una nueva intención que contenga esta opción con su respectivo plan asociado. Las opciones estarán representadas por los planes aplicables asociados. En G-JASON se modificó el método "selectora de opciones" ("*selectOption*") para que seleccione la opción con mayor relevancia (planRelevance). Luego, si se presentan las opciones:

$P1 = @Plan1[planRelevance(0.85)]+sensor\_termico(TEMP)[degOfCert(0.7)] : (TEMP < 300)[degOfCert(0.6)] < -encender\_alarma; llamar\_supervisor,$

$P2 = @Plan2[planRelevance(0.65)]+sensor\_termico(TEMP)[degOfCert(0.5)] : (TEMP < 100)[degOfCert(0.3)] < -encender\_calentador; encender\_turbina,$

La selectora retorna la opción P1 debido a que es la opción (plan) con mayor relevancia (planRelevance).

Selección de intenciones: en la selección de opciones del proceso se crearon o actualizaron el conjunto de intenciones que posee el agente. El conjunto de intenciones será la entrada del método de selección de intenciones, donde se eligirá una intención a ejecutar para terminar el ciclo de interpretación del agente. Cada intención estará representada por el plan que cada una de ellas contiene y el cuerpo de ese plan contiene una lista de acciones para su posterior ejecución. El orden de ejecución de acciones de una intención se mantiene según la versión original de JASON. Al igual que en la selección de opciones, la selectora de intenciones tendrá en cuenta el valor del grado en los planes (*planRelevance*) y se seleccionará la intención con mayor grado de relevancia (planRelevance).Una vez seleccionada la intención, la máquina de estados ejecuta una de las acciones contenidas en el cuerpo del plan.

5 Caso de Estudio

Se presenta un caso de estudio que muestra el potencial de las extensiones anteriormente descriptas. Se desea modelar la supervisión de un horno rotativo utilizado para la fundición de metales. Se analizan constantemente tres variables fundamentales en la operación: temperatura, presión y vibración. Para esta tarea se cuenta

con tres sensores estrategicamente situados. Por lo tanto, los sensores proveen al agente de información sobre las lecturas y sus grados de certeza asociados los cuales dependerán de la precisión de los instrumentos utilizados. El agente además posee acciones recomendadas de acuerdo a las lecturas de los sensores, las cuales serán modeladas dentro de los planes. Para el caso de la variable presión por ejemplo, si la lectura supera los 35 bars se deberá cerrar una válvula de inyección del horno (plan P3) y si la presión supera los 70 bars se deberá encender una alarma general, dado el peligro de explosión en el horno y se deberán tomar medidas de precaución (plan P4).

Se sabe además, que la precisión de los sensores de temperatura son de 70% a 700 grados por lo tanto el grado de certeza de la lectura de temperatura (B1) será de 0.7. Para los planes se agregarán grados (planRelevance) para dar relevancia a la necesidad de urgencia de ejecutar ciertas acciones por el agente supervisor. Dada la relevancia de la presión para los hornos, se agregan los grados de relevancia más altos a los planes relacionados a la presión P3 (0,9) y P4 (0,85) respectivamente.

B1= $sensor\_termico(700)[degOfCert(0.7)]$

B2= $sensor\_presion(80)[degOfCert(0.9]$.

B3= $sensor\_vibracion(8)[degOfCert(0.6)]$.

P1= $@alerta\_temperatura[planRelevance(0.7)]$

$+sensor\_termico(TEMP)[degOfCert(0.5)]$:

$TEMP > 300[degOfCert(0.6)] < -prende\_ventilador$.

P2= $@urgencia\_temperatura[planRelevance(0.8)]$

$+sensor\_termico(TEMP)[degOfCert(0.5)]$:

$TEMP > 600[degOfCert(0.6)] < -$**apagar_horno**.

P3= $@alerta\_presion[planRelevance(0.85)]+sensor\_presion(PRES)[degOfCert(0.7)]$:

$PRES > 35[degOfCert(0.8)] < -cierra\_valvula$.

P4= $@urgencia\_presion[planRelevance(0.9)]+sensor\_presion(PRES)[degOfCert(0.7)]$:

$PRES > 70[degOfCert(0.8)] < -$**enciende_alarma**.

P5= $@manejo\_vibracion[planRelevance(0.6)]+sensor\_vibracion(NVL)[degOfCert(0.4)]$:

$NVL > 5[degOfCert(0.5)] < -$**frena_motor**.

Sin considerar un archivo de prioridades, el orden de las acciones resultantes a ejecutar por el agente son: **(1) enciende_alarma, (2) apaga_horno y (3)frena_motor**.

Se observa que la acción *enciende_alarma* es la primera en ejecutarse. Esta acción esta relacionada con el manejo de presión modelada por el belief B2 (grado de certeza 0,9) el cual es el más confiable. Además, de los planes aplicables (P3 y P4) se ha ejecutado primero el plan P4 (relevancia de 0,9), por ser el plan más relevante. Ejecutando este ejemplo en la versión de JASON original, no hay posibilidad de explicitar la certeza o prioridad, en las creencias o en los planes y se ejecutó primero la acción relacionada a la temperatura dado el orden de llegada de las lecturas de los sensores. El agente supervisor implementado en G-JASON mejora su performance al atender primero las lecturas más confiables (sensor de presión) y ejecutar las acciones vinculadas a este evento, en el orden más relevante.

6 Conclusiones

Se ha realizado una extensión de JASON tanto sintáctica como semántica para incluir grados en las creencias y en los planes (en distintos componentes). También se ha agregado el concepto de prioridades para poder ordenar los eventos según el orden de reactividad que se les quiera otorgar. Al incluir las prioridades y grados mencionados, para que los mismos tengan el valor semántico deseado, se tuvieron que modificar las funciones selectoras y los procesos de unificación para los eventos y los contextos de los planes. Se implementaron todas las modificaciones necesarias, en una nueva versión de JASON que denominamos G-JASON. A través de un caso de estudio se pudo comprobar que con esta extensión se aumenta la expresividad de JASON pudiendo representar una situación que involucra incertidumbre en los hechos y diferentes relevancias de planes obteniendo mejores resultados que en JASON original. Respecto al modelo BDI graduado que ha inspirado este trabajo ha quedado pendiente modelar grados en los deseos (goals de JASON) ya que en la estructura de los sistemas PRS estos elementos no son considerados en los componentes básicos y no tienen un representación adecuada. Se plantea como trabajo futuro, trabajar en la modelización de los grados de deseos del agente en estas plataformas.

References

1. Bordini, R., Hübner, J.: BDI Agent Programming in AgentSpeak Using JASON. John Wiley and Sons. (2007)
2. Casali, A., Godo, Ll., Sierra, C.: A graded BDI agent model to represent and reason about preferences: Artificial Intelligence, Special Issue on Preferences Artificial Intelligence. vol. 175, pp. 1468–1478. (may, 2011)
3. Casali, A., Godo, Ll., Sierra, C.: Lecture Notes in Artificial Intelligence: Graded BDI Models For Agent Architectures. Leite, Joao and Torroni, Paolo, Springer-Verlag. 126–143. Berling Heidelberg (2005)
4. Casali, A., Godo, Ll., Sierra, C.: A Tourism Recommender Agent: From theory to practice. In: Revista Iberoamericana de Inteligencia Articial, vol. 12:40, pp. 23–38. (2008)
5. D'Inverno, M., Kinny, D., Luck, M., Wooldridge, M.: A formal specification of dMARS. Intelligent Agents IV: Proc. Fourth International Workshop on Agent Theories, Architectures and Languages. Singh, M.P. and Rao, A.S. and Wooldridge, M. Springer-Verlag, 155–176, Montreal, (1998)
6. Georgeff, M. P., Lansky, A. L.: Reactive reasoning and planning. AAAI-87, 677–682, Seattle. (1987)
7. Georgeff, M., Pell, B., Pollack, M., Tambe, M., Wooldridge, M.: The Belief-Desire-Intention Model of Agency. Intelligent Agents. Muller, J. P. and Singh, M. and Rao, A. S. Springer-Verlag, vol. 1365, Berling Heidelberg. (1999)
8. Krapf, A., Casali, A.: Desarrollo de Sistemas Inteligentes aplicados a redes elctricas industriales. In: WASI-CACIC, Corrientes, Argentina (2007)
9. Rao, A., Georgeff, M.: BDI Agents from Theory to Practice. In: AAII. (1995)
10. Wooldridge, M.: Introduction to Multiagent Systems, 2 Ed., John Wiley and Sons, (2009).

ABN: Considerando Características de los Objetos de Negociación

Pablo Pilotti[1], Ana Casali[1,2] and Carlos Chesñevar[3]

[1] Centro Internacional Franco Argentino de
Ciencias de la Información y de Sistemas (CIFASIS) Rosario,
Av. 27 de febrero 210 bis - S2000EZP, Rosario, Argentina
Email: pilotti@cifasis-conicet.gov.ar

[2] Facultad de Cs. Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario (UNR)
Email: acasali@fceia.unr.edu.ar

[3] Depto. de Cs. e Ingeniería de la Computación
Universidad Nacional del Sur (UNS) - CONICET
Email: cic@cs.uns.edu.ar

Resumen Un escenario típico de negociación involucra dos agentes cooperativos que poseen recursos y tienen la necesidad de realizar un intercambio para alcanzar sus objetivos. Los recursos tienen características que pueden ser determinantes para alcanzar o no, sus metas. Los agentes conocen con certeza tanto sus recursos como las características que estos tienen, pero suelen tener creencias, posiblemente erróneas o incompletas, sobre los recursos del oponente. Para acordar cuales recursos trocar, establecen una comunicación en donde se ofrecen propuestas de posibles intercambios y se responden con críticas y contrapropuestas. El rumbo de la comunicación va cambiando en la medida de que los agentes revelan los recursos que poseen y las características de estos. En este trabajo se presenta un modelo de negociación automática basada en argumentación para agentes que quieren negociar en este tipo de escenarios. Se utilizan técnicas de revisión de creencias para la interpretación de las propuestas y argumentación para la generación de propuestas.

1. Introducción

La negociación es una forma de interacción entre dos o más agentes que tienen objetivos en conflicto y buscan un acuerdo aceptable. Un escenario típico de negociación involucra a dos agentes co-dependientes (necesitan del otro agente para alcanzar su objetivo) que deben negociar para alcanzar un beneficio mutuo. Diferentes enfoques se han desarrollado para abordar la negociación en sistemas multiagentes [7] entre las cuales se puede destacar la negociación basada en argumentación –*ABN: argumentation-based negotiation*– (por ej. [1], [4], [6],[7],). En ABN se plantea que los agentes que negocian no sólo intercambien propuestas, sino que también las razones que soportan a las mismas.

Esto permite a los agentes conocer las creencias y preferencias de la contraparte modificando la aceptabilidad y valoración de cada propuesta [3]. Además, dado que los agentes generalmente tienen información incorrecta o incompleta sobre los otros, el intercambio de argumentos les brinda información que les posibilita actualizar sus creencias.

Para participar en el proceso de negociación, los agentes ABN tienen que ser capaces de generar, seleccionar y evaluar los argumentos asociados a las propuestas y actualizar acordemente sus estados mentales[6]. En [5] los autores han presentado un modelo de negociación automática entre dos agentes cooperativos, benevolentes (i.e. siempre tratan de hacer lo que se les pide) y veraces (i.e. no comunican información que conocen que es falsa). Las propuestas son modeladas como argumentos, se consideran recursos tanto a los objetos físicos como al conocimiento necesario para alcanzar los objetivos y se utilizan técnicas de revisión de creencias para el proceso de interpretación y generación de argumentos. En ese modelo se permite intercambiar conjuntos de recursos, sin embargo no se considera la descripción de características que tienen los objetos y por lo tanto, en el diálogo no se incluyen argumentos que traten sobre estas características.

Las características de los objetos son importantes en el proceso de una negociación en el contexto planteado, ya que determinan si se puede con estos recursos alcanzar o no un objetivo. Por ejemplo, generalmente con un martillo y un clavo se puede colgar un cuadro en la pared, pero si el clavo es pequeño suele no servir para ese fin. En [1] los autores muestran como a través de un diálogo que incluye propiedades de los objetos/sujetos a negociar en los argumentos, puede influir en las preferencias de los agentes que negocian y por ende, en el resultado de la negociación.

En este trabajo se extiende el modelo de negociación basada en argumentación presentado en [5] en distintos aspectos. En primer lugar, se incluye la representación de las características de los objetos y su utilización en los argumentos que soportan o critican a las distintas propuestas. Luego, se extiende las formalización de una propuesta para que incluya la argumentación de lo que se ofrece. Además, los mensajes que intercambian los agentes son más completos, pudiendo incluir una crítica a la última propuesta recibida. Para dar una mejor representación al conocimiento que tiene un agente, sobre los planes que le permitan alcanzar o no, los distintos objetivos (los cuales dependen de las propiedades de los recursos que están involucrados), se ha decidido utilizar una lógica rebatible para su representación.

Se considera un escenario de negociación que involucra a dos agentes que poseen recursos y tienen la necesidad de realizar un intercambio (trueque) para alcanzar sus objetivos. Los recursos tienen características que pueden ser determinantes para alcanzar o no, sus metas. Los agentes conocen con certeza sus objetivos, sus recursos y las propiedades que estos tienen y por otra parte, tienen creencias, posiblemente incompletas o erróneas, sobre los objetivos y los recursos del oponente. Además, pueden tener conocimiento erróneo o incompleto sobre cuáles son los planes y recursos que le permiten satisfacer sus propias metas.

Para acordar cuáles recursos trocar, establecen una comunicación en donde se intercambian mensajes. Dentro de estos mensajes se ofrecen posibles intercambios (propuestas) y se responden con críticas y contra-ofertas. En la medida de que los agentes revelan los recursos que poseen y las características que estos tienen, las creencias de los agentes respecto al oponente van cambiando y con ello, se modifican las propuestas a intercambiar. En este trabajo se formaliza un modelo de negociación automática basada en argumentación para este tipo de escenario, utilizando técnicas de revisión de creencias para la interpretación y la generación de las propuestas. A continuación se presenta un ejemplo motivador del modelo de negociación que se propone.

Example 1. Dos agentes Ag_1 y Ag_2 se encuentran en una habitación, cada uno tiene un objetivo diferente que no pueden alcanzar con sus propios recursos.

El agente Ag_1 tiene como objetivo decorar una pared. Posee los siguientes recursos: tiene un cuadro pequeño, un pincel, un tornillo y un martillo. Cree que: colgando un cuadro a la pared o pintándola de rojo (solamente) la pared quedará decorada, usando un martillo y un clavo usualmente se puede colgar el cuadro, usando un martillo y un clavo pequeño usualmente se puede colgar un cuadro pequeño y que usando un tornillo y un destornillador usualmente se puede colgar un espejo. Respecto al agente Ag_2 cree que su objetivo es pintar la pared y que posee un destornillador y pintura pero no de color rojo.

Por el otro lado, el agente Ag_2 tiene como objetivo colgar un espejo en la pared. Tiene los siguientes recursos: un espejo, una tachuela y un destornillador. Tiene las siguientes creencias: una tachuela es un clavo chico, con un martillo y un clavo usualmente se puede colgar un cuadro, pero si el clavo es chico usualmente no se lo puede colgar, con un martillo y un clavo usualmente se puede colgar un espejo, pintar la pared la decora. También tiene como creencia, que el objetivo del agente Ag_1 es pintar una pared.

Adicionalmente ambos creen: que si la pared está pintada de rojo, entonces está pintada y que con un pincel y pintura usualmente se pinta la pared,

Cabe destacar que en este escenario, los agentes se encuentran en la siguiente situación: (1) ninguno de los dos puede alcanzar su meta por si mismo, (2) tienen información incompleta o incorrecta respecto del otro agente (por ej. la creencia de Ag_2 sobre el objetivo de Ag_1 es incorrecta) (3) tienen información contradictoria respecto a como alcanzar un objetivo (por ej. si se puede o no, colgar un cuadro utilizando un clavo pequeño).

El resto de este trabajo está estructurado de la siguiente forma, en la Sección 2 se presenta el lenguaje que utiliza un agente y en la Sección 3 se describe su arquitectura. Luego, en la Sección 4 se introduce el diálogo entre agentes y la Sección 5 los procesos de evaluación, interpretación y generación de las propuestas. Finalmente, se presenta en la Sección 6 algunas conclusiones y en el apéndice un diálogo posible entre los agentes del ejemplo motivador, .

2. Representación del conocimiento del Agente

En este trabajo se utiliza Defeasible Logic Programing (DeLP) como lenguaje de representación y mecanismo de deducción de los agentes, que permite tratar

conocimiento rebatible. Luego las creencias de cada uno son representadas como un "Defeasible Logic Program". A continuación se introduce DeLP de manera compacta, el lector puede referirsea [2] para una presentación completa de DeLP. En DeLP, un programa P es un par (Π, Δ) donde Π es un conjunto consistente de hechos y reglas estrictas y Δ un conjunto de reglas rebatibles. Los hechos son representados por literales (átomos o átomos negados que utilizan la negación fuerte "$\sim$"), las reglas estrictas se denotan "$l_0 \leftarrow l_1, .., l_n$" mientras que las rebatibles "$l_0 \prec l_1, .., l_n$". Las reglas rebatibles representan información tentativa, pueden ser utilizadas para la deducción de la conclusión si no se plantea nada en su contra, mientras que los hechos y reglas estrictas representan el conocimiento no rebatible. Por lo tanto, una regla rebatible representa una conexión débil entre el cuerpo y la cabeza, y debe leerse como "si $l_1, .., l_n$ entonces usualmente se cumple l_0'. En DeLP también existe la posibilidad de representar hechos negados y reglas cuyas cabezas contienen literales negados. Por lo tanto es posible la derivación de conclusiones contradictorias. Un conjunto $A \subset \Delta$ es un *argumento* para el literal l (denotado $\langle A, l \rangle$) si $\Pi \cup A$ es consistente y deriva a l. Para establecer si $\langle A, l \rangle$ es no–derrotado, se lo compara con todos los contra–argumentos es decir argumentos que lo refutan o contradicen y por algún criterio son preferidos a $\langle A, l \rangle$. Dado que los contra–argumentos también son argumentos, estos pueden ser derrotados por otros, y así formar una secuencia de argumentos. Una *línea de argumentación aceptable* es una secuencia finita de argumentos $[A_1, .., A_n]$ tal que: Cada argumento derrota a su predecesor; son consistentes tanto la union de los argumentos que soportan a la conclusión (A_{2k+1}) como la unión de los que la que desafían (A_{2k}); no existen argumentos que sean sub argumentos de otros que aparezca antes en la secuencia. Por lo general, cada argumento tiene más de un argumento que lo derrota y por lo tanto, podría existir más de una línea de argumentación. El conjunto de líneas de argumentación forma un árbol, llamado árbol dialéctico, que tiene argumentos como nodos y cada camino desde la raíz a una hoja representa una línea de argumentación. El árbol se utiliza para decidir si un literal l se justifica o no. Si en un árbol dialéctico cuya raíz es $\langle A, l \rangle$ existe una línea de argumentación de longitud par, significa que existe un argumento que desafía a la conclusión y que no puede ser derrotado. En estos casos el árbol dialéctico no justifica al literal l. Un literal l es justificado en un programa DeLP, si existe un árbol dialéctico que lo justifica. De esta manera, dado un programa P, y una consulta l el intérprete DeLP responde: *YES* si P justifica l, *NO* si P justifica $\sim l$, *UNDECIDED* si P no justifica ni l ni $\sim l$, *UNKNOWN* si no pertenece al lenguaje del programa.

En el contexto de negociación establecido, los agentes negocian objetos que poseen características con el fin de satisfacer sus metas. De esta manera, los objetos, las características y las metas son *elementos de la negociación* que deben ser representados por el lenguaje y es necesario distinguirlos en el lenguaje DeLP,

Los literales que se distinguen para representar los elementos de la negociación son los siguientes: $\mathcal{R}^{at}$, un conjunto de fórmulas atómicas que representan clases de recursos (ej. $hammer(h)$, $nail(n)$, $picture(p)$), $\mathcal{F}^{at}$, un conjunto de fórmulas atómicas que representan características de recursos (ej. $red(c)$,

$small(n)$), $\mathcal{G}^{at}$ un conjunto de fórmulas atómicas que representan metas (ej. $decorate\_wall$, $hang(p)$).

Los conjuntos de reglas que se distinguen para representar la relación entre los elementos de la negociación son los siguientes:

$\mathcal{R}_d = \{r_0 \prec r_1, f_1, .., f_n \mid r_i \in \mathcal{R}^{at}, f_j \in \mathcal{F}^{at}\}$. El conjunto de las fórmulas en forma de regla que relacionan recursos y características con otros recursos Estas reglas representan la creencia de que un recurso r_0 usualmente es considerado un recurso r_1 con las características $f_1, ..., f_n$. Por ejemplo la fórmula $tack(c) \prec nail(c), small(c)$ indica que una tachuela es un clavo chico, y la fórmula $\sim tack(c) \prec nail(c), \sim small(c)$ indica que si el clavo no es chico entonces no es considerada tachuela;

$\mathcal{R}_p = \{f_0 \prec r_0, \mid r_0 \in \mathcal{R}^{at}, f_0 \in \mathcal{F}^{at}\}$, el conjunto de las fórmulas en forma de regla que relacionan recursos con características, estas reglas representan la creencia de que un recurso r_0 usualmente tiene la característica f_0. Por ejemplo la fórmula $small(c) \prec tack(c)$ indica que una tachuela es considerada como algo chico;

$\mathcal{G}_d = \{g_0 \leftarrow g_1, .., g_n \mid g_i \in \mathcal{G}^{at}\}$, el conjunto de las fórmulas en forma de regla que relacionan metas con otras metas. Estas reglas significan que la meta g_0 es alcanzada si todos las sub-metas $g_1, ..g_n$ son alcanzadas. Por ejemplo: $decorate\_wall \leftarrow red\_paintwork(c)$ significa que una pared se decora si se pinta de rojo;

$\mathcal{P} = \{g \prec r_1, f_1^1, .., f_1^m, .., r_n, f_n^1, .., f_n^l \mid g \in \mathcal{G}^{at}, r_i \in \mathcal{R}^{at}, f_i^j \in \mathcal{F}^{at}\}$ el conjunto de las fórmulas en forma de regla que relacionan las metas con los objetos y características. Este tipo de regla significa que la meta g usualmente se puede alcanzar con el recurso r_1 que posee las características $f_1^1, .., f_1^m$, con el recurso r_2 con características $f_2^1, .., f_2^k$ y así sucesivamente. Por ejemplo, la fórmula $hang(p) \prec picture(p), hammer(h), nail(n)$ significa que un cuadro se puede colgar utilizando un martillo y un clavo, mientras que la fórmula $\sim hang(p) \prec picture(p), hammer(h), nail(n)$ sugiere que no es posible si el clavo es pequeño.

De esta manera, las fórmulas del lenguaje que describen los recursos se denotan como $\mathcal{R} = \mathcal{R}^{at} \cup \mathcal{F}^{at} \cup \mathcal{R}_d \cup \mathcal{R}_p$ y las que describen las metas $\mathcal{G} = \mathcal{G}^{at} \cup \mathcal{G}_d$.

En este trabajo las creencias de cada agente están enfocadas a representar los recursos que posee y su descripción ($\mathcal{R}$), las metas y las sub-metas necesarias para alcanzar otras metas ($\mathcal{G}$) y los recursos que son necesarios para alcanzar metas ($\mathcal{P}$). Se define al lenguaje de creencias como: $\mathcal{B} = \mathcal{R} \cup \mathcal{G} \cup \mathcal{P}$.

3. Arquitectura del Agente

El estado mental de cada agente negociador Ag_i posee un conjunto de creencias (B_i), un objetivo a alcanzar (G_i), así como también un conjunto de creencias (B_i^j) sobre su agente oponente Ag_j y cual es el objetivo de su oponente (G_i^j). Además de las creencias, el estado mental de los agentes incluye el historial de negociación, es decir el diálogo establecido desde el comienzo de la negociación (Ver Def. 4 para mas detalles), lo cual se formaliza a continuación.

Definition 1. *Sean Ag_i, Ag_j los agentes involucrados en la negociación. El* ***estado mental*** *del agente Ag_i y de manera equivalente el de Ag_j es una tupla $MS_i = < B_i, G_i, B_i^j, G_i^j, H_i >$, donde: $B_i, B_i^j \subset \mathcal{B}$; $G_i, G_i^j \in \mathcal{G}^{at}$ y H es la historia de negociación entre Ag_i y Ag_j.*

Continuando con el ejemplo 1, se formaliza a Ag_1 como
$MS_1 =< B_1, G_1, B_1^2, G_1^2, H_1 >$, donde:

$B_1 = \{picture(p), small(p), brush(b), screw(s), hammer(h),$
$decorate\_wall \leftarrow hang(p);\ \ decorate\_wall \leftarrow red\_paintwork(c)$
$\sim decorate\_wall \leftarrow\sim red\_paintwork(c)$
$hang(p) \prec picture(p), hammer(h), nail(n)$
$hang(p) \prec picture(p), small(p), hammer(h), nail(n), small(n)$
$hang(m) \prec mirror(m), screw(s), screwdriver(sc)$
$paintwork(c) \prec brush(b), paint(c);\ \ paintwork(c) \prec red\_paintwork(c)\ \}$
$G_1^1 = decorate\_wall$
$B_1^2 = \{paint(c), \sim red(c), screwdriver(sc)\}$
$G_1^2 = paintwork(c)$
$H_1 = \emptyset$

El mecanismo de decisión de los agentes utiliza la información del estado mental (MS) para calcular que mensaje enviar al otro agente. La función $Init$ calcula el primer mensaje del diálogo, los demás mensajes son calculados por $Answer$.

Definition 2. *El* ***mecanismo de decisión*** *de un agente Ag_i consiste en una tupla $DM_i =< Init_i, Answer_i >$, donde:*

$Init_i : MS_i \rightarrow MS_i \times Message$
$Answer_i : MS_i \times Message \rightarrow MS_i \times Message \cup \{accept, withdraw\}$

El algoritmo $Init$ es el encargado de comenzar la negociación, en primer lugar selecciona las propuestas que cree que sirven tanto para alcanzar su meta como la del otro agente, si no existen selecciona las propuestas que solamente sirven para alcanzar su meta sin tener en cuenta si lo ofrecido le es útil al otro agente, si tampoco existen envía el mensaje $withdraw$ indicando que abandona la negociación. Por otro lado, el algoritmo $Answer$ actualiza las creencias del agente utilizando la propuesta recibida (proceso de interpretación). A continuación, si lo ofrecido le permite alcanzar su meta y puede cumplir con lo que le demandan, envía el mensaje *accept* indicando que está de acuerdo (proceso de evaluación). En caso contrario genera una crítica y una contra propuesta (proceso de generación). Los procesos de interpretación, evaluación y generación de propuestas se detallan en la sección 5.

Finalmente se define un agente negociador como un estado mental, que lleva cuenta de la información relevante para la negociación y por un mecanismo de decisión que calcula el próximo mensaje del diálogo. En la próxima sección se definen las propuestas y críticas que pueden ser enviadas en los mensajes.

Definition 3. *Un* ***agente*** *Ag_i es una tupla $< MS_i, DM_i >$, donde MS_i representa su estado mental y DM_i su mecanismo de decisión.*

4. Propuestas y Críticas para Alcanzar un Acuerdo

El diálogo entre dos agentes esta determinado por una secuencia finita de mensajes enviados de manera alternada por cada uno de los agentes participantes de la negociación. Los mensajes contienen propuestas y críticas, el mensaje final del diálogo es uno que indica que hay acuerdo entre los agentes (*accept*) o bien uno que indica que no hay acuerdo (*withdraw*).

Definition 4. *Un **diálogo** entre dos agentes negociadores Ag_i y Ag_j es una secuencia finita de mensajes $[m_1, ..., m_{n-1}, m_n]$ tal que: (1) m_1 es una propuesta p_1, (2) para $1 < i < n$, m_i es un par (c_i, p_i) donde c_i es una crítica a p_{i-1} y p_i es una propuesta. (3) $m_n \in \{accept, withdraw\}$, (4) no hay mensajes repetidos, i.e. $m_s \neq m_t$, with $t, s < n$; (5) dos mensajes consecutivos no pertenecen al mismo agente.*

Una propuesta es una sentencia que deja al descubierto la intención que tiene un agente de intercambiar un conjunto de recursos que posee (Oferta) por otro conjunto de recursos que necesita (Demanda). Adicionalmente la propuesta incluye una justificación de porque esta interesado en realizar ese intercambio. Una propuesta puede verse como la siguiente sentencia:

`Necesito` B_d`, porque` B_{ds}`, luego alcanzo` G_d`.`
`A cambio te doy` B_o`, porque` B_{os}`, luego alcanzás` G_o`.`

donde B_d representa el conjunto de recursos que el agente demanda, B_{ds} representa el conjunto de creencias que utiliza el agente para justificar que con los recursos demandados alcanza la meta G_d. De manera análoga, B_o representa el conjunto de recursos que el agente ofrece a cambio, B_{os} representa el conjunto de creencias que el agente utiliza para justificar que con los recursos ofrecidos se alcanza la meta G_o. Estos conceptos son formalizados en la Definición 5.

Definition 5. *Sean B_d, B_{ds}, B_o, B_{os} subconjuntos de $B_{\mathcal{L}}$; y G_d, G_o elementos de $\mathcal{G}^{at}$. Una **Propuesta** es una tupla $\langle d, o\rangle$, donde el primer elemento es una demanda $d = \langle B_d, B_{ds}, G_d\rangle$, el segundo una oferta $o = \langle B_o, B_{os}, G_o\rangle$ y se cumple lo siguiente:*[4]

1. $B_d \cup B_{ds} \mid\sim G_d$ *2.* $B_{ds} \mid\not\sim G_d$ *3.* $B_o \cup B_{os} \mid\sim G_o$ *4.* $B_{os} \mid\not\sim G_o$

Continuando con el ejemplo 1, suponga que Ag_1 comienzala negociación con el siguiente mensaje:
Necesito un clavo porque con el martillo que tengo, podría amurar un cuadro a la pared y de esta manera quedaría decorada la pared. Quiero decorar la pared. A cambio te doy un pincel para que lo uses con la pintura que tenés para pintar la pared.

Formalmente este mensaje se representa como $p_1 = \langle d_1, o_1\rangle$, donde:

$$d_1 = \langle \{nail(n)\}, \{hammer(h); picture(p); \\ hang(p) \prec picture(p), hammer(h), nail(n); decorate\_wall \leftarrow hang(p)\}, \\ decorate\_wall\rangle$$
$$o_1 = \langle \{brush(b)\}, \{paint(c); paintwork(c) \prec brush(b), paint(c)\}, paintwork(c)\rangle$$

[4] Se escribe $P \mid\sim a$ cuando la respuesta de la consulta a en el programa DeLP P es YES y $P \mid\not\sim a$ cuando la respuesta es NO o UNDECIDED.

Luego de que un agente recibe una propuesta, él puede contestar con un mensaje: de aceptación (*accept*), en cuyo caso se realiza el intercambio de recursos de acuerdo a la última propuesta; de salida de la negociación (*withdraw*) en este caso no se efectúa ningún intercambio; o bien, puede rechazar la propuesta sin salir de la negociación haciendo una crítica a la última propuesta y agregar una nueva propuesta.

Definition 6. *Sea una propuesta* $p = \langle d, o\rangle$*, donde:* $d = \langle B_d, B_{ds}, G_d\rangle$ *y* $o = \langle B_o, B_{os}, G_o\rangle$*, se define una* ***crítica*** *c como* $\langle (C_1, C_2), C_3\rangle$ *donde* $C_1 \subset D_{\mathcal{L}}$*,* $C_1 \subset B_{\mathcal{L}}$*,* $C_3 \in \{\emptyset, yes, no\}$ *y alguna de las siguientes condiciones se cumple:*

1. $(B_d - C_1) \cup B_{ds} \not\sim G_d$
2. $B_d \cup B_{ds} \cup C_2 \not\sim G_d$
3. $C_3 = no$

Se observa que C_1 y C_2 son críticas a la demanda (d_1), básicamente la primer crítica afirma la falta de los recursos demandados, la segunda crítica adiciona información de manera que la meta no pueda ser justificada. También se puede observar que C_3 es una crítica a la oferta (o_1), básicamente es un rechazo a la oferta, sin especificar la causa. Continuando con el ejemplo, Ag_2 podría responder a la propuesta p_1 usando alguna de las siguientes críticas:

$C_1 = \{nail(n)\}$ *(No tengo un clavo.)*

$C_2 = \{tack(n);\ nail(c) \prec tack(c);\ small(c) \prec tack(c);$
$\sim hang(p) \prec picture(p), hammer(h), nail(n), small(n)\}$
(Tengo una tachuela, que es un clavo chico. Si el clavo es chico, usualmente no se puede amurar un cuadro.)

$C_3 = no$ *(No creo que tu oferta sea útil para mi.)*

5. Interpretación, Evaluación y Generación de Propuestas

El proceso de interpretación comienza luego de que un agente recibe una propuesta, éste está basado en las siguientes intuiciones: Como los agentes son veraces, benevolentes y conocen sus propios recursos, cuando el agente Ag_j recibe una propuesta $p = \langle\langle B_d, B_{ds}, G_d\rangle, \langle B_o, B_{os}, G_o\rangle\rangle$ de Ag_i, el primero puede inferir lo siguiente: (1) que Ag_i no tiene B_d (lo que demanda) y si tiene B_o (lo que ofrece), (2) que Ag_i cree en B_{ds} y B_{os}, (3) que Ag_i tiene como objetivo G_d. A partir de estas inferencias, el estado mental se actualiza utilizando operadores de revisión de creencias, de manera similar [5].

A continuación el agente realiza un proceso de evaluación de la propuesta. Dado que un agente es consciente de los recursos, creencias y objetivos que posee, puede determinar de manera individual si: (1) posee los recursos demandados, (2) no existe ningún conjunto de creencias que refuten la demanda, (3) que la oferta de la propuesta permita alcanzar la meta. En caso de que todas estás condiciones se cumplan, el agente acepta la propuesta, por el contrario si alguna condición no se cumple entonces puede generar la crítica adecuada.

De manera similar a [5], el proceso de generación de propuestas Gen se basa en un operador $\oplus$ que calcula todas las pruebas tentativas de un programa DeLP. A continuación se define el operador $\oplus$ y la función Gen de la siguiente manera:

Definition 7. *Sea P un programa DeLP, y h un literal, se define el conjunto $P \oplus h$ de la siguiente manera: $X \in P \oplus h$ sii X es la unión de hechos y reglas que participan en algún árbol dialéctico que garantiza h en el programa P.*

Podemos ver que cada elemento $X \in P \oplus h$, (1) es un programa DeLP, (2) garantiza h, (3) si se elimina un hecho o regla de los argumentos que soportan la conclusión, entonces X deja de garantizar h.

Definition 8. *Sean $B', B'' \subset \mathcal{B}$ y $G' \subset \mathcal{G}$, se define* Gen *de la siguiente manera:*
$\mathsf{Gen}(B', B'', G') =_{def} \{ \langle B, B_s, G\rangle : \ B, B_s, \subset \mathcal{B},\ G \subset \mathcal{G},\ \ G = G'$
$(B_s \cup B) \in (B' \cup R'' \cup B) \oplus G',\ \}$

Proposition 1. *Sea un agente con estado mental $MS_i =< B_i, G_i, B_i^j, G_i^j, H_i >$ se cumple lo siguiente: Si $d \in \mathsf{Gen}(B_i, B_i^j, G_i)$, $o \in \mathsf{Gen}(B_i^j, B_i, G_i^j)$ y la oferta (o) y demanda (d) no comparten recursos entonces: $\langle o, d\rangle$ es una propuesta*

6. Conclusiones

En este trabajo se ha presentado un modelo de negociación automática basada en argumentación donde los agentes utilizan conocimiento rebatible y técnicas de revisión de creencias para la interpretación y la generación de las propuestas. Se ha extendido el modelo de negociación basada en argumentación presentado en [5] en distintos aspectos. En primer lugar, se incluyó la representación de las características de los objetos y su utilización en los argumentos que soportan o critican a las distintas propuestas. También, se extendió las formalización de una propuesta para que incluya la argumentación de lo que se ofrece. Además, los mensajes que intercambian los agentes son más complejos, pudiendo incluir una crítica a la última propuesta recibida. Para dar una mejor representación al conocimiento que tiene cada agente, sobre los planes que le permitan alcanzar o no, los distintos objetivos (los cuales dependen de las propiedades de los recursos que están involucrados), se utilizó una lógica rebatible para su representación.

Actualmente se trabaja en la extensión y adaptación de la implementación de los agentes desarrollados en Prolog en [5] para poder implementar y experimentar con agentes dentro de este modelo más amplio de negociación.

Apéndice: Diálogo entre Agentes

A continuación se presenta un diálogo posible entre los dos agentes del ejemplo motivador:

Ag_1 *Me das un clavo? porque quiero decorar la pared. Con el martillo que tengo, podría amurar un cuadro a la pared y de esta manera quedaría decorada la pared. A cambio te doy un pincel para que lo uses con la pintura que tenés para pintar la pared.*

$d = \langle \{nail(n)\}, \{hammer(h); picture(p); decorate\_wall \leftarrow hang(p);$
$hang(p) \prec picture(p), hammer(h), nail(n);\}, decorate\_wall\rangle$

$o = \langle \{brush(b)\}, \{paint(c); paintwork(c) \prec brush(b), paint(c)\}, paintwork(c)\rangle$

Ag_2 *Tengo una tachuela, que es un clavo chico, y si el clavo es chico,usualmente el cuadro no se puede amurar a la pared.*

$C_2 = \{tack(n);\ nail(c) \prec tack(c);\ small(c) \prec tack(c);$
$\sim hang(p) \prec picture(p), hammer(h), nail(n), small(n)\}$
Me das un martillo? porque quiero colgar un espejo. Con la tachuela que tengo, podría amurar el espejo. A cambio te doy pintura para que lo uses con el pincel que tenés para pintar la pared, creo que si la pared esta pintada entonces está decorada, por lo tanto decorás la pared.

$d = \langle \{hammer(h)\}, \{tack(n); mirror(m); nail(n) \prec tack(n)$
$hang(m) \prec mirror(m), hammer(h), nail(n); \}, hang(m) \rangle$

$o = \langle \{paint(c)\}, \{brush(b); paintwork(c) \prec brush(b), paint(c);$
$decorate\_wall \leftarrow paintwork(c)\}, decorate\_wall \rangle$

Ag_1 *La pintura no es de color rojo, por lo tanto no la decora.*

$C_2 = \{\sim red(p); \sim red\_paintwork(c) \prec brush(b), paint(p), \sim red(p)$
$\sim decorate\_wall \leftarrow \sim red\_paintwork(c), paintwork(c)\}$
Me das un clavo chico? porque quiero decorar la pared. Con el martillo que tengo, podría amurar un cuadro a la pared porque este es pequeño. y de esta manera quedaría decorada la pared. A cambio te doy un tornillo para que lo uses con el destornillador y cuelgues el cuadro.

$d = \langle \{nail(n), small(n)\}, \{hammer(h); picture(p);$
$hang(p) \prec picture(p), small(p), hammer(h), nail(n), small(n);$
$decorate\_wall \leftarrow hang(p)\}, decorate\_wall \rangle$

$o = \langle \{screw(s)\}, \{hang(m) \prec mirror(m), screw(s), screwdriver(sc);$
$screwdriver(sc)\}, hang(m) \rangle$

Ag_2 *Acepto*

Referencias

1. L. Amgoud and S. Vesic. A formal analysis of the outcomes of argumentation-based negotiations. In Liz Sonenberg, Peter Stone, Kagan Tumer, and Pinar Yolum, editors, *AAMAS*, pages 1237–1238. IFAAMAS, 2011.
2. Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: an argumentative approach. *Theory Pract. Log. Program.*, 4(2):95–138, January 2004.
3. N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra, and M. Wooldridge. Automated negotiation: Prospects, methods and challenges. *International Journal of Group Decision and Negotiation*, 10(2):199–215, 2001.
4. S. Parsons, C. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292, 1998.
5. Pablo Pilotti, Ana Casali, and Carlos Chesñevar. A belief revision approach for argumentation-based negotiation with cooperative agents. In *9th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2012)*, Valencia, Spain, 2012.
6. I. Rahwan, P. Pasquier, L. Sonenberg, and F. Dignum. On the benefits of exploiting underlying goals in argument-based negotiation. In *Twenty-Second Conference on Artificial Intelligence (AAAI)*, pages 116–121, Vancouver, 2007.
7. I. Rahwan, S. D. Ramchurn, N. R. Jennings, P. Mcburney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *Knowl. Eng. Rev.*, 18:343–375, December 2003.

Dinámica de Conocimiento: Contracción Múltiple en Lenguajes Horn

Néstor Jorge Valdez[1] **Marcelo A. Falappa**[2]

[1] Departamento de Ciencias de la Computación, Fac. de Ciencias Exactas y Naturales
Universidad Nacional de Catamarca (UNCa)
Av. Belgrano 300 - San Fernando del Valle de Catamarca
Tel.: (03834)420900 / Cel: (03834) 154591186
e-mail: njvaldez@c.exactas.unca.edu.ar
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Av. Alem 1253, (B8000CPB) Bahía Blanca, Argentina
Tel: (0291)4595135 / Fax: (0291)4595136
e-mail: mfalappa@cs.uns.edu.ar

Abstract. En los últimos años el estudio de la teoría de cambio de creencias dentro del marco AGM, ha motivado la necesidad de desarrollar modelos de la teoría de contracción que abarquen los casos de contracción simultánea para conjuntos de sentencias y no solamente para una única sentencia. Por ello, en este paper se presentan algunos modelos que resultan ser generalizaciones de funciones de contracción AGM, pero considerando el caso de las contracciones de un conjunto de sentencias, especialmente bajo el fragmento Horn de la lógica proposicional. Además, se consideró que las definiciones de los distintos modelos de contracción Horn obtenidas basadas en las contracciones múltiples, resulten tan plausible como una contracción AGM. También, se demuestra que las contracciones Horn obtenidas satisfacen este criterio establecido, como así también se proporcionan las pruebas que identifican los postulados que la caracterizan.

1. Introducción

1.1. Motivación

La teoría de cambio de creencias estudia la forma en que un agente cambia sus creencias cuando adquiere nueva información. Los cambios implican, a menudo, la eliminación de creencias existentes (operación de contracción) e incorporación de creencias adquiridas (operación de revisión) [8]. El modelo dominante de cambio de creencias es conocido como modelo AGM [1], el cual lleva ese nombre por las iniciales de sus tres creadores: Carlos Alchourrón, Peter Gärdenfors y David Makinson. El modelo AGM para cambio de creencias asume una lógica subyacente que es al menos tan expresiva como la lógica proposicional. Debido a este supuesto, esta teoría no puede ser aplicada a los sistemas con lógicas subyacentes que son menos expresivos que la lógica proposicional. Para este paper nos enfocamos en el estudio de las operaciones de cambio múltiple, i.e., procedimientos de cambio de creencias que se llevan a cabo simultaneamente para un conjunto de sentencias [11]. La investigación se centra principalmente en los modelos constructivos más conocidos de la contracción de creencias (*partial*

meet contractions [1] y *kernel contraction* [15]) en el cual la lógica subyacente se basa en un fragmento de la lógica proposicional, que es la lógica Horn. Recordemos que una cláusula de Horn es una disyunción de literales que consisten de, a lo sumo, un literal positivo, e.g. $\neg p \vee \neg q \vee s$ y decimos que una lógica Horn está constituído por conjunciones de cláusulas Horn. La contracción Horn ha sido materia de estudio en [3, 4, 6], ello se debe a su amplio campo de aplicación tanto en la Inteligencia Artificial, Bases de Datos, así como en Ontologías en Lógicas Descriptivas [2].

1.2. Preliminares

Consideramos un lenguaje proposicional $\mathcal{L}$, sobre un conjunto de literales $\mathbf{P} = \{p, q, \ldots\}$, con semánticas de un modelo teórico estándar. Los caracteres griegos en minúsculas $\varphi, \psi, \ldots$ denotan fórmulas y los caracteres en mayúsculas $X, Y, \ldots$ denotan conjuntos de fórmulas. Una cláusula Horn es una cláusula con a lo sumo un literal positivo. Una fórmula Horn es una conjunción de cláusulas Horn. Una teoría Horn es un conjunto de fórmulas Horn. El lenguaje Horn $\mathcal{L}_H$ es una restricción de $\mathcal{L}$ para fórmulas Horn. La lógica Horn obtenida de $\mathcal{L}_H$ tiene la misma semántica que la lógica proposicional obtenida de $\mathcal{L}$, pero restringida para fórmulas Horn y sus derivables. La consecuencia lógica clásica y su equivalencia lógica se denota por $\vdash$ y $\equiv$ respectivamente. Cn es el operador de consecuencia tal que $Cn(X) = \{\varphi \mid X \vdash \varphi\}$. La consecuencia lógica bajo lógica Horn se denota por $\vdash_H$ y así el operador de consecuencia Cn_H bajo lógica Horn es tal que $Cn_H(X) = \{\varphi \mid X \vdash_H \varphi\}$. Los conjuntos clausurados se representarán mediante letras en negrita. Por ejemplo, si $\mathbf{K}$ es un conjunto de creencias entonces $\mathbf{K} = Cn(\mathbf{K})$.

El presente artículo está organizado de la siguiente manera: en la sección 2 se presenta algunos conceptos y definiciones con respecto a una extensión del marco AGM y que ha motivado la necesidad de desarrollar modelos de contracciones que consideren el caso de contracción por un conjunto de sentencias (simultáneas) y no por una sola sentencia, en la sección 3 se recuerda brevemente algunos conceptos fundamentales y necesarios sobre cambio de creencias bajo lógica Horn que juegan un rol importante para el desarrollo y presentación de resultados en este trabajo de investigación, en la sección 4 se ofrecen parte de las contribuciones de este artículo donde se proporcionan una definición apropiada según las notaciones incorporadas hasta aquí, como asi tambien una representación de resultados del tipo de contracción múltiple denominada *package contraction* para lógicas Horn, por último en la sección 5 están las conclusiones y futuras investigaciones.

2. Contracción Múltiple: una extensión del modelo AGM

Contraer por un conjunto de sentencias en lugar de por una sola fue presentado por [10], quien uso el término contracción múltiple para designar este tipo de operaciones. Otros autores que también estudiaron la teoría de las operaciones de cambio múltiple fueron [11, 12, 16, 17]. Fermé, Saez y Sanz [9] ampliaron el campo de conocimiento presentando dos maneras de generalización de las funciones de contracción kernel para conjuntos de sentencias (no necesariamente clausurados), y sobre conjuntos de creencias (clausurados). Una contracción múltiple de un conjunto de creencias $\mathbf{K}$ por un conjunto de sentencias B significa la eliminación del conjunto B de $\mathbf{K}$. Podemos también interpretar esta idea de las siguientes maneras:

- La eliminación de todos los elementos de B de $\mathbf{K}$. Es decir, que el resultado de $K \div [B]$ de la contracción múltiple de K por B debe ser tal que $B \cap (\mathbf{K} \div [B]) = \varnothing$.
- La eliminación de al menos uno de los elementos de B de $\mathbf{K}$. Es decir, que el resultado de $\mathbf{K} \div \langle B \rangle$, de la contracción múltiple de $\mathbf{K}$ por B debe ser tal que $B \nsubseteq \mathbf{K} \div \langle B \rangle$.

Fuhrmann y Hansson [11], denominan a la primera clase de contracción múltiple descripta anteriormente descrita como *package contraction* y a las de segunda clase como *choice contraction*. En la misma investigación ellos presentan para conjuntos de creencias, dos operaciones de la primera clase y una de la segunda clase. Así para *package contraction* sugieren las operaciones *partial meet package contraction* y *subremainder contraction*.

Los potenciales resultados de la contracción por paquetes de una teoría $\mathbf{K}$ por un conjunto de sentencias, por ejemplo $\{\alpha, \beta\}$ pueden en general ser diferentes de cada conjunto, pudiendo tener como resultado cualquiera de las siguientes operaciones:

1. Contraer $\mathbf{K}$ por $\alpha \wedge \beta$,
2. Contraer $\mathbf{K}$ por $\alpha \vee \beta$,
3. Contraer primero por α y luego contraer el resultado de tal contracción por β, o al revés,
4. Intersectar los resultados de contraer $\mathbf{K}$ por α y de contraer $\mathbf{K}$ por β.

Algunas observaciones que se obtienen con respecto a las operaciones mencionadas anteriormente se detallan en [11]. Allí, se demuestra formalmente que el resultado de la partial meet package contraction de $\mathbf{K}$ por un conjunto B no resulta ser idéntica al conjunto que resulta de la intersección de los resultados de contraer $\mathbf{K}$ por cada uno de las sentencias en B. Aquí, consideraremos esencialmente las contracciones múltiples de las clases de *package* en el contexto de la modelación *partial meet* para conjuntos finitos.

A continuación presentamos un conjunto de postulados que constituyen las propiedades intuitivamente necesarias en una función de contracción múltiple, en [9, 11, 12, 13, 14] se pueden encontrar algunas interrelaciones entre sus postulados. Asumiremos que $\mathbf{K}$ es un conjunto de creencias y B, C son conjuntos arbitrarios de sentencias.

- **Package Closure:** $\mathbf{K} \div B$ es un conjunto de creencias (i.e. $\mathbf{K} \div B = Cn(\mathbf{K} \div B)$).
- **Package Inclusion:** $\mathbf{K} \div B \subseteq \mathbf{K}$.
- **Package Vacuity:** Si $B \cap \mathbf{K} = \varnothing$, entonces $\mathbf{K} \div B = \mathbf{K}$.
- **Package Success:** Si $B \cap Cn(\varnothing) = \varnothing$, entonces $B \cap \mathbf{K} \div B = \varnothing$.
- **Package Extensionality:** Si para cada sentencia α en B existe una sentencia β en C tal que $\vdash \alpha \leftrightarrow \beta$, y vice versa, entonces $\mathbf{K} \div B = \mathbf{K} \div C$.
- **Package Recovery:** $\mathbf{K} \subseteq Cn((\mathbf{K} \div B) \cup B)$.
- **Finite Package Recovery:** Si B es finito, entonces $\mathbf{K} \subseteq Cn((\mathbf{K} \div B) \cup B)$.
- **Package Uniformity:** Si cada subconjunto X de $\mathbf{K}$ implica algún elemento de B si y solamente si X implica algún elemento de C, entonces $\mathbf{K} \div B = \mathbf{K} \div C$.
- **Package Relevance:** Si $\beta \in \mathbf{K}$ y $B \notin \mathbf{K} \div B$, entonces existe un conjunto K' tal que $\mathbf{K} \div B \subseteq K' \subseteq \mathbf{K}$ y $B \cap Cn(K') = \varnothing$ pero $B \cap Cn(K' \cup \{\beta\}) \neq \varnothing$.
- **Package Core-Retainment:** Si $\beta \in \mathbf{K}$ y $B \notin \mathbf{K} \div B$, entonces existe un conjunto K', tal que $K' \subseteq \mathbf{K}$ y $B \cap Cn(K') = \varnothing$ pero $B \cap Cn(K' \cup \{\beta\}) \neq \varnothing$.

2.1. Partial Meet Multiple Contraction

Teniendo en mente los conceptos básicos de funciones partial meet contraction referido a una única sentencia [1], presentaremos los conceptos fundamentales de las *partial meet multiple contractions*.

Definición 1 [11, 12] *Sea* $\mathbf{K}$ *un conjunto de creencia,* B *un conjunto de sentencias y* $\mathbf{K} \perp B$ *el conjunto de restos de* K *con respecto a* B*. Una* package selection function *para* $\mathbf{K}$ *es una función* γ *tal que para todos los conjuntos de sentencias* B*:*

1. *Si* $\mathbf{K} \perp B$ *es no-vacío, entonces* $\gamma(\mathbf{K} \perp B)$ *es un subconjunto no vacío de* $\mathbf{K} \perp B$*, y*
2. *Si* $\mathbf{K} \perp B$ *es vacío, entonces* $\gamma(\mathbf{K} \perp B) = \{\mathbf{K}\}$.

Entonces, la definición de *partial meet multiple contraction* producto de la generalización de *partial meet contraction* para el caso de contracciones por conjuntos de sentencias es:

Definición 2 *(**Partial meet multiple contraction** [11, 12] Sea* $\mathbf{K}$ *un conjunto de sentencias y sea* γ *una package selection function para* $\mathbf{K}$*. La partial meet multiple contraction de* $\mathbf{K}$ *generada por* γ *es la operación* $\div_\gamma$ *tal que para cualquier conjunto de sentencias de* B*:*

$$\mathbf{K} \div_\gamma B = \bigcap \gamma(\mathbf{K} \perp B)$$

Una multiple contraction function $\div$ *de* $\mathbf{K}$ *es una partial meet multiple contraction si y solamente si existe alguna package selection function* γ *tal que* $\mathbf{K} \div B = \mathbf{K} \div_\gamma B$ *para cualquier conjunto de sentencias* B.

Las definiciones de los dos casos limites particulares de *partial meet contractions* son:

Definición 3 *Sea* $\mathbf{K}$ *un conjunto de creencias.*

1. *Una multiple contraction function* $\div$ *en* $\mathbf{K}$ *es una maxichoice multiple contraction si y solamente si es una partial meet multiple contraction generado por un package selection function* γ *tal que para todos los conjuntos* B*, el conjunto* $\gamma(\mathbf{K} \perp B)$ *tiene exactamente un elemento.*
2. *La full meet multiple contraction en* $\mathbf{K}$ *es el partial meet multiple contraction* $\doteqdot$ *que es generado por la package selection function* γ *tal que para todos los conjuntos* B*, si* $\mathbf{K} \perp B$ *es no-vacío, entonces* $\gamma(\mathbf{K} \perp B) = \mathbf{K} \perp B$*, i.e., la multiple full meet contraction* $\doteqdot$ *es la operación de contracción en* $\mathbf{K}$ *definido por:*

$$\mathbf{K} \doteqdot B = \begin{cases} \bigcap \mathbf{K} \perp B & \text{si } B \bigcap Cn(\varnothing) = \varnothing \\ \mathbf{K} & \text{en caso contrario} \end{cases}$$

para cualquier conjunto B.

2.2. Kernel Multiple Contraction

Se presenta a continuación la definición de *contracción múltiple kernel* que resulta ser una generalización de la operación de contracción kernel para una sola sentencia, pero que se refiere para contracciones por conjuntos de sentencias [1].

Definición 4 [9] *Sea* A *y* B *dos conjuntos de sentencias. El package kernel set de* A *con respecto a* B*, denotado* $A \Subset_P B$ *es el conjunto tal que* $X \in A \Subset_P B$ *si y solamente si:*

1. $X \subseteq A$.
2. $B \cap Cn(X) \neq \varnothing$.
3. *Si* $Y \subset X$ *entonces* $B \cap Cn(Y) = \varnothing$.

Esta definición es más general pues A no necesariamente es un conjunto de creencias. La definición de package incision function para un conjunto A, que resulta en una función que selecciona al menos un elemento de cada uno de los conjuntos en $A \Subset_P B$, para cualquier conjunto B.

[1] Fuhrmann y Hansson definen las multiple partial meet contraction sobre conjuntos de creencias o belief set. Ferme y otros definen las multiple kernel contraction sobre conjuntos de sentencias (conjuntos arbitrarios, no necesariamente clausurados)

Definición 5 [9] *Una función σ es una función de incisión para A si y solamente si, para cualquier B:*

1. $\sigma(A \perp\!\!\!\perp_P B) \subseteq \bigcup A \perp\!\!\!\perp_P B$.
2. *Si* $\varnothing \neq X \in A \perp\!\!\!\perp_P B$, *entonces* $X \cap \sigma(A \perp\!\!\!\perp_P B) \neq \varnothing$.

Definición 6 *(**Kernel Multiple Contraction** [**9**]) Sea σ una incision function para A. La kernel multiple contraction $\approx_\sigma$ para A basado en σ esta definida como sigue:*

$$A \approx_\sigma B = A \backslash \sigma(A \perp\!\!\!\perp_P B).$$

Una multiple contraction function $\div$ para A es una kernel multiple contraction si y solamente si existe alguna package incision fuction σ para A tal que $A \div B = A \approx_\sigma B$ para cualquier B.

3. Contracción de Creencias Horn

Delgrande presentó los primeros resultados sobre cambio de creencias Horn [6], investigando la analogía Horn entre *orderly maxichoice contraction* y las *orderly maxichoice Horn contraction*, las cuales están basadas en la noción de *remainder set*. En [6] presentan la definición bajo fragmento Horn. La representación de resultados para OMHC es la siguiente:

Teorema 1 [6] *Sea $\dot{-}$ una función de contracción Horn. Para cada conjunto de creencias Horn H, $\dot{-}$ es una* orderly maxichoice Horn contraction function *si y solo si satisface:*

$(H \dot{-} 1)$ $H \dot{-} \varphi = Cn_H(H \dot{-} \varphi)$ *(closure)*

$(H \dot{-} 2)$ $H \dot{-} \varphi \subseteq H$ *(inclusion)*

$(H \dot{-} 3)$ *Si* $\varphi \notin H$, *entonces* $H \dot{-} \varphi = H$ *(vacuity)*

$(H \dot{-} 4)$ *Si* $\nvdash \varphi$, *entonces* $\varphi \notin H \dot{-} \varphi$ *(success)*

$(H \dot{-} 6)$ *Si* $\varphi \equiv \psi$, *entonces* $H \dot{-} \varphi = H \dot{-} \psi$ *(extensionality)*

$(H \dot{-} f)$ *Si* $\vdash \varphi$, *entonces* $H \dot{-} \varphi = H$ *(failure)*

$(H \dot{-} ce)$ *Si* $\psi \notin H \dot{-} \varphi \wedge \psi$, *entonces* $H \dot{-} \varphi \wedge \psi = H \dot{-} \varphi$ *(conjunctive equality)*

Booth, Meyer et al. [3] presentaron la *infra Horn contraction* IHC como variante de PMC que satisface la así llamado *convexity property*. Ella establece que cualquier conjunto de creencias que es un subconjunto del conjunto de creencias que se obtiene por *maxichoice contraction* y un superconjunto que es obtenido por un *full meet contraction* puede ser obtenido por algunas PMCs.

Teorema 2 *Sea K un conjunto de creencias. Sea $\dot{-}_{mc}$ un* maxichoice contraction *para K y $\dot{-}_{fm}$ la* full meet contraction *para K. Para cada $\varphi \in \mathcal{L}$ y cada conjunto de creencia X tal que $K \dot{-}_{fm} \varphi \subseteq X \subseteq K \dot{-}_{mc} \varphi$, existe un* partial meet contraction $\dot{-}_{pm}$ *para K tal que $K \dot{-}_{pm} \varphi = X$.*

En [5] se define la *infra contraction* $\dot{-}$ para K. Para la construcción de una *infra Horn contraction* el interés es preservar la propiedad de convexidad para poder dar todas las contracciones Horn apropiadas. Para la versión Horn las adaptaciones de las definiciones expresadas anteriormente para una infra contraction son las siguientes:

Definición 7 [5] *Sea H un conjunto de creencias Horn y φ una fórmula. El conjunto de* infra remainder sets *de H con respecto a φ, denotado como $H \downarrow_i \varphi$, es tal que $X \in H \downarrow_i \varphi$ si y solo si existe un $Y \in H \downarrow_m \varphi$, siendo $Y \in H \downarrow_m \varphi$ el conjunto de* maxichoice remainder sets, *tal que*

$$X = Cn(X) \text{ y } (\bigcap H \downarrow_m \varphi) \subseteq X \subseteq Y.$$

Definición 8 [5] *Sea H un conjunto de creencias Horn y τ una* infra selection function *para H. Una* infra Horn contraction *$\div$ para H, que esta determinado por τ, es tal que para toda fórmula φ:*

$$H \div \varphi = \tau(H \downarrow_i \varphi)$$

Proposición 1 *También es posible demostrar que la infra contraction es idéntica a partial meet contraction y kernel contraction bajo lógica Horn.*

Definición 9 [5] *Sea H un conjunto de creencias Horn y φ una fórmula Horn. El conjunto de* kernel sets *de H con respecto a φ, denotado como $H \Downarrow \varphi$, es tal que $X \in H \Downarrow \varphi$ si y solo si*

1. *$X \subseteq H$*
2. *$X \vdash \varphi$, y*
3. *Si $Y \subset X$, entonces $Y \nvdash \varphi$.*

Los elementos de $H \Downarrow \varphi$ son los $\varphi - kernels$ de H.

Definición 10 [5] *Sea H un conjunto de creencias Horn y σ una función de incisión para H. Una* kernel Horn contraction *$\div$ para H, que esta determinado por σ, es tal que:*

$$H \div \varphi = Cn_H(H \setminus \sigma(H \Downarrow \varphi))$$

para todo $\varphi \in \mathcal{L}_H$.

Otra de las variantes Horn de partial meet contraction es la partial meet Horn contraction. Estas contracciones para ser válidas deben permitir una exacta correspondencia con PMC. Delgrande y Wassermann [7] introdujeron la construcción de PMHC también basado en la noción de Horn remainder set y al que denominaron *weak remainder set*. Se pretende con la definición de weak remainder set preservar las propiedades del modelo teórico del conjunto de restos estándar. De esta manera, se conserva la correspondencia entre PMHC y PMC. Se llega a esta conclusión debido a la relación entre conjuntos de restos y su contrapartida en término de interpretaciones. Delgrande demostró que los modelos de un conjunto de resto consiste de los modelos de un conjunto H de creencias agregado a ello un contramodelo de la fórmula φ para contracción. Pero esto no ocurre generalmente con cláusulas Horn, donde para un contramodelo M de φ, es posible que no encontremos un conjunto de resto Horn que tenga a M como un modelo. Como resultados propusieron los llamados *weak remainder sets*. Algunas de sus definiciones y caracterizaciones:

Definición 11 [7] *Sea H un conjunto de creencias Horn, φ una fórmula Horn y m un modelo del conjunto de modelos de un conjunto H de creencias. $H \Downarrow_w \varphi$ es el conjunto de conjuntos de fórmulas tal que $H' \in H \Downarrow_w \varphi$ si y solo si $H' = H \cap m$ para algún $m \in \mid \neg\varphi \mid$. $H' \in H \Downarrow_w \varphi$ es un* weak remainder set *de H y φ.*

Definición 12 [7] *Sea H un conjunto de creencias Horn. γ es una función de selección para H si, para cada $\varphi \in \mathcal{L}_H$,*

1. *Si $H \Downarrow_w \varphi \neq \varnothing$ entonces $\varnothing \neq \gamma(H \Downarrow_w \varphi) \subseteq H \Downarrow_w \varphi$.*
2. *Si $H \Downarrow_w \varphi = \varnothing$ entonces $\gamma(H \Downarrow_w \varphi) = \{H\}$.*

En [7] definen el *weak remainder set* y su función de selección para H. Entonces, la construcción de PMHC es equivalente a la PMC con la variante que en lugar de usar los conjuntos de restos estándar se recurre a los conjuntos de restos débiles.

Definición 13 [7] *Sea H un conjunto de creencias Horn y γ una función de selección para H. Una* partial meet Horn contraction *$\div$ para H, que está determinado por γ, es tal que:*

$$H \div_w \varphi = \bigcap \gamma(H \Downarrow_w \varphi)$$

y si $\gamma(H \Downarrow_w \varphi) = \{H'\}$ para algún $H' \in H \Downarrow_w \varphi$ su maxichoice Horn contraction basado en weak remainders estaría dado por:

$$H \div_w \varphi = \gamma(H \Downarrow_w \varphi)$$

para todo $\varphi \in \mathcal{L}_H$.

4. Contracción Múltiple de Creencias Horn

Ahora, como parte de las contribuciones de este artículo proporcionamos una definición apropiada según las notaciones incorporadas hasta aquí, como así también una representación de resultados del tipo de contracción múltiple denominada *package contraction* pero para lógicas Horn. El procedimiento de remover un conjunto de sentencias de un conjunto de creencias H es contraer con la disyunción de las sentencias a eliminar en la lógica proposicional clásica. Con la lógica Horn esto se complica, debido a que no considera las disyunciones totales o completas (sentencias compuesta solamente por disyunciones). Para formalizar la operación de contraer un conjunto de sentencias Φ con respecto a un conjunto H con fragmento Horn consideraremos los conjuntos de restos. Considerar la lógica Horn como la lógica subyacente con respecto a la contracción AGM clásica nos permitirá adaptar la contracción de conjuntos finitos de sentencias Φ. Delgrande [6] demostró que es posible realizar este movimiento para obtener otros tipos de contracciones (*entailment-based contraction y inconsistency-based contraction*). Por lo tanto, el comportamiento en una e-contraction con respecto a un conjunto de sentencias Φ es el mismo con respecto a una sola sentencia. A continuación realizamos las adaptaciones de las diferentes definiciones de una e-contraction para obtener las respectivas para p-contraction.

Definición 14 *Sea H un conjunto de creencias Horn y Φ un conjunto de fórmulas Horn. Decimos que $H' \in H \downarrow_p \Phi$ si y solo si*

1. *$H' \subseteq H$,*
2. *$Cn(H') \cap \Phi = \varnothing$,*
3. *Para todo H'' tal que $H' \subset H'' \subseteq H$, $Cn(H'') \cap \Phi \neq \varnothing$.*

y llamamos los Horn p-remainder sets de H con respecto a Φ a los elementos de $H \downarrow_p \Phi$.

La definición de las *partial meet Horn p-selection functions* es:

Definición 15 *Una partial meet Horn p-selection functions σ es una función de $\mathcal{P}(\mathcal{P}(\mathcal{L}_H))$ a $\mathcal{P}(\mathcal{P}(\mathcal{L}_H))$ tal que*

1. *$\sigma(H \downarrow_p \Phi) = \{H\}$ si $H \downarrow_p \Phi = \varnothing$,*
2. *$Y \varnothing \neq \sigma(H \downarrow_p \Phi) \subseteq H \downarrow_p \Phi$ en otro caso.*

Ahora estamos en condiciones de establecer la de *partial meet Horn p-contraction.*

Definición 16 *Dado una partial meet Horn p-selection function σ, $\div_\sigma$ es una partial meet Horn p-contraction si y solo si*

$$H \div_\sigma \Phi = \bigcap \sigma(H \downarrow_p \Phi).$$

Para la definición de los dos casos extremos *maxichoice y full meet Horn p-contraction* es como sigue.

Definición 17 *Dado una partial meet Horn p-selection function σ, $\div_\sigma$ es un maxichoice Horn p-contraction si y solo si:*

$$\sigma(H \downarrow_p \Phi) \text{ es un conjunto simple o conjunto unitario.}$$

Análogamente, $\div_\sigma$ es un full meet Horn p-contraction si y solo si:

$$\sigma(H \downarrow_p \Phi) = H \downarrow_p \Phi \text{ cuando } H \downarrow_p \Phi \neq \varnothing.$$

Ahora, de la misma manera podemos trabajar con los *infra p-remainder sets* y obtener una definición formal para *Horn p-contraction.*

Definición 18 *Sea H un conjunto de creencias Horn y Φ un conjunto de fórmulas Horn. Decimos que $H' \in H \Downarrow_p \Phi$ si y solo si*

existe algún $H'' \in H \downarrow_p \Phi$ *tal que* $(\bigcap H \downarrow_p \Phi) \subseteq H' \subseteq H''$

y llamamos los infra p-remainder sets de H *con respecto a* Φ *a los elementos de* $H \Downarrow_p \Phi$.

Obtenemos ahora una definición para *Horn p-contraction* en términos de *infra p-remainder sets*.

Definición 19 *Una infra p-selection functions* τ *es una función de* $\mathcal{P}(\mathcal{P}(\mathcal{L}_H))$ *a* $\mathcal{P}(\mathcal{P}(\mathcal{L}_H))$ *tal que*

1. $\tau(H \Downarrow_p \Phi) = H$ *cuando* Φ *es tautológico,*
2. $\tau(H \Downarrow_p \Phi) \in H \Downarrow_p \Phi$ *en otro caso.*

Una función de contracción $\div_\tau$ *es una Horn p-contraction si y solo si* $H \div_\tau \Phi = \tau(H \Downarrow_p \Phi)$.

La representación de resultados para Horn p-contraction resulta sencillo ya que también se realizan las adaptaciones de los postulados correspondientes. Los postulados que caracterizan la Horn p-contraction son:

$(H \div_p 1)$ $H \div_p \Phi = Cn(H \div_p \Phi)$

$(H \div_p 2)$ $H \div_p \Phi \subseteq H$

$(H \div_p 3)$ Si $H \cap \Phi = \varnothing$ entonces $H \div_p \Phi = H$

$(H \div_p 4)$ Si Φ no es tautológico entonces $(H \div_p \Phi) \cap \Phi = \varnothing$

$(H \div_p 5)$ Si $\Phi \equiv \Psi$ entonces $H \div_p \Phi = H \div_p \Psi$

$(H \div_p 6)$ Si $\varphi \in H \setminus (H \div_p \Phi)$, existe un H' tal que $\bigcap(H \downarrow_p \Phi) \subseteq H' \subseteq H$, $Cn(H') \cap \Phi = \varnothing$, y $(H' + \varphi) \cap \Phi \neq \varnothing$

$(H \div_p 7)$ Si Φ es tautológico entonces $H \div_p \Phi = H$

Por último, definimos *Horn package contraction* y su relación con *maxichoice Horn contraction*, todo ello basado en *weak remainders*. Empezamos realizando la adaptación de la definición **11** para conjunto de sentencias Φ.

Definición 20 *Sea* H *un conjunto de creencias Horn y* Φ *un conjunto de fórmulas Horn.* $H \Downarrow_p \Phi$ *es el conjunto de conjuntos de fórmulas tal que* $H' \in H \Downarrow_p \Phi$ *si y solo si*

1. $H' \subseteq H$,
2. *para cada* $\varphi \in \Phi$ *donde* $\varphi \notin Cn_H(\top)$, $H' \subseteq m$ *para algún* $m \in |\neg\varphi|$,
3. *para cada* H'' *donde* $H' \subset H'' \subseteq H$, *tenemos* $H'' \nsubseteq m$ *para algún* $\varphi \in \Phi$ *donde* $m \in |\neg\varphi|$.

Adaptamos la definición **12** para denotar su función de selección para conjunto de sentencias Φ.

Definición 21 *Sea* H *un conjunto de creencias Horn.* γ *es una función de selección para* H *tal que* $\gamma(H \Downarrow_p \Phi) = \{H'\}$ *para algún* $H' \in H \Downarrow_p \Phi$.

Obtenemos de la definición **13** la *package Horn contraction* basado en *weak remainders*.

Definición 22 *Sea* H *un conjunto de creencias Horn y* γ *una función de selección para* H, *la (maxichoice) package Horn contraction basado en weak remainders está dado por*

$$H \div_p \Phi = \gamma(H \Downarrow_p \Phi)$$

si $\varnothing \neq \Phi \cap H \nsubseteq Cn_H(\top)$, *y* H *en otro caso.*

Mediante el siguiente teorema Delgrande y Wassermann [7] establece que cualquier maxichoice Horn contraction define una package contraction, y vice versa.

Teorema 3 [7] *Sea H un conjunto de creencias Horn y sea $\Phi = \{\varphi_1, ..., \varphi_n\} \subset \mathcal{L}_H$. Tenemos que $H' \in H \Downarrow_p \Phi$ si y solo si $H' = \bigcap_{i=1}^{n} H_i$ donde $H_i \in H \Downarrow_e \varphi_i$, $1 \leqslant i \leqslant n$.*

Considerando la proposición **1**, el teorema 8 de [5], el teorema 7 de [4], como así también lo demostrado por Falappa [8] en el contexto base de creencias, es posible generalizar lo establecido en infra contraction y kernel contraction con una sentencia para su generalización con conjunto de sentencias bajo fragmento Horn. No ocurre lo mismo si las operaciones están basadas en weak remainder, esto se debe a las diferencias técnicas entre ellas. Entre las principales diferencias podemos mencionar por ejemplo, que los weak remainder y los e-remainder son conceptos distintos, ya que la operación partial meet corresponde a las intersecciones de weak remainder. Otra diferencia relevante consiste en que *no todos los infra remainders son weak remainders* como así también *no todos los weak remainders son infra remainders*. Booth et al [3], demuestran mediante ejemplos como en algunos casos el conjunto que se obtiene es un infra remainder pero no un weak remainder y como en otros si coinciden. Además, los infra remainders deben contener un full meet y estar contenidos en algún remainder (por definición). En cambio, los weak remainders están contenidos en algún remainder o ser un remainder, pero no siempre contienen un full meet contraction (como lo demuestran Booth et al).

5. Conclusión y Trabajos Futuros

Al realizar la adaptación de las operaciones de contracción bajo lógica Horn por sentencias simples a su generalización para un conjunto de sentencias hemos realizado una importante contribución para la investigación dentro de la contracción para lógica Horn. En resumen, las principales contribuciones del presente paper son:

i) generalización de las operaciones de contracción bajo lógica Horn partial meet e infra basado en remainder set con sentencia simple a sus correspondientes package Horn contraction, y considerando que para la maxichoice package Horn contraction su representación de resultado se logra sustituyendo el postulado de weak relevance por un postulado de maximalidad.

ii) a partir de resultados de investigaciones previas se demuestran que las operaciones infra contraction coinciden con partial meet contraction.

iii) la demostración de que no es posible la generalización de las operaciones apuntadas en el item ii), cuando éstas se basan en weak remainder sets.

Para este artículo, nos enfocamos en la caracterización de las operaciones de contracción Horn múltiples: partial meet (y sus variantes basados en infra y weak remainder) y kernel. Como trabajo futuro se planea extender la investigación para lograr su generalización y representación de resultados de contracciones múltiples a otras operaciones de contracción conocidas del framework AGM como *epistemic entrenchment* bajo lógica Horn, entre otras.

Referencias

[1] Carlos Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50:510–530, 1985.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. The description logic handbook. *CUP, Cambridge*, 2003.

[3] Richar Booth, Thomas Meyer, and Iván José Varzinczak. Next steps in propositional horn contraction. *In Boutilier, C. (Ed.), Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 702–707, 2009.

[4] Richard Booth, Thomas Meyer, Iván José Varzinczak, and Renata Wassermann. A contraction core for horn belief change: Preliminary report. *In 13th International Workshop on Nonmonotonic Reasoning (NMR), (2010a/b)*, 2010.

[5] Richard Booth, Thomas Meyer, Iván José Varzinczak, and Renata Wassermann. On the link between partial meet, kernel, and infra contraction and its application to horn logic. *Journal of Artificial Intelligence Research*, pages 31–53, 2011.

[6] James P. Delgrande. Horn clause belief change: Contraction functions. *In Gerhard Brewka and Jérome Lang, editors, Proceedings of the Eleventh International Conference on the Principles of Knowledge Representation and Reasoning, Sydney, Australia, 2008. AAAI Press*, pages 156–165, 2008.

[7] James P. Delgrande and Renata Wassermann. Horn clause contraction functions: Belief set and belief base approaches. In Fangzhen Lin, Ulrike Sattler, and Miroslaw Truszczynski, editors, *KR*. AAAI Press, 2010.

[8] Marcelo A. Falappa, Eduardo L. Fermé, and Gabriele Kern-Isberner. On the logic of theory change: Relations between incision and selection functions. *In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, Proceedings of the 17th European Conference on Artificial Intelligence, ECAI 2006*, pages 402–406, 2006.

[9] Eduardo Fermé, Karina Saez, and Pablo Sanz. Multiple kernel contraction. *Studia Logica*, 73:183–195, 2003.

[10] André Fuhrmann. Relevant logics, modal logics and theory change. *PhD thesis, Australian National University, Camberra*, 1988.

[11] Andre Fuhrmann and Sven Ove Hansson. A survey of multiple contractions. *Journal of Logic, Language and Information*, pages 39–76, 1994.

[12] Sven Ove Hansson. New operators for theory change. *Theoria*, 55:114–132, 1989.

[13] Sven Ove Hansson. Belief base dynamics. *PhD thesis, Uppsala University*, 1991.

[14] Sven Ove Hansson. Belief contraction without recovery. *Studia Logica 50(2)*, pages 251–260, 1991.

[15] Sven Ove Hansson. Kernel contraction. *J. of Symbolic Logic 59(3)*, pages 845–859, 1994.

[16] Reinhard Niederée. Multiple contraction: A further case against gärdenfors' principle of recovery. *In A. Fuhrmann and M. Morreau, editors, The Logic of Theory Change. Berlin, 1991. Springer-Verlag*, pages 322–334, 1991.

[17] Hans Rott. Modellings for belief change: Base contraction, multiple contraction, and epistemic entrenchment (preliminary report). *In D. Pearce and G. Wagner, editors, Logics in AI. Springer Berlin / Heidelberg*, 633:139–153, 1992.

Intelligent Algorithms for Reducing Query Propagation in Thematic P2P Search

Ana Lucía Nicolini, Carlos M. Lorenzetti,
Ana G. Maguitman, and Carlos Iván Chesñevar

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur, Av. Alem 1253, (8000) Bahía Blanca, Argentina
{aln,cml,agm,cic}@cs.uns.edu.ar

Abstract. Information retrieval is a relevant topic in our days, especially in distributed systems where thousands of participants store and share large amounts of information with other users. The analysis, development and testing of adaptive search algorithms is a key avenue to exploit the capacity of P2P systems to lead to the emergence of semantic communities that are the result of the interaction between participants. In particular, intelligent algorithms for neighbor selection should lead to the emergence of efficient communication patterns. This paper presents new algorithms which are specifically aimed at reducing query propagation overload through learning peers' interests. Promising results were obtained through different experiments designed to test the reduction of query propagation when performing thematic search in a distributed environment.

1 Introduction

The current information age has facilitated the generation, publication and access to geographically dispersed resources and heterogeneous content. As searching and sharing information directly from personal computers become more prevalent, new opportunities arise to preserve, foster and exploit the diversity of social communities in Internet. In this scenario we can identify several research challenges for developing mechanisms to manage and access distributed resources in a variety of formats. While research on peer-to-peer (P2P) systems has facilitated the implementation of robust distributed architectures, there are still several limitations faced by current search mechanisms. In particular, these mechanisms are unable to reflect a thematic context in a search request and to effectively take advantage of the peers' interests to improve the network communication patterns.

The main objective of this work is to provide P2P systems with mechanisms for context-based search and to propose algorithms that incrementally learn effective communication patterns in pure P2P networks, where each participant operates in an autonomous manner, without relying on a specific server for communications.

Current search services are rigid as they do not offer mechanisms to facilitate users access to information about potentially relevant topics with which they might not be familiar. Another limitation of the current search model is the lack of context sensitivity. Although some websites offer personalized search they do not offer proper mechanisms to facilitate contextualization and collaboration. These factors are crucial in thematic and distributed search environments.

In a distributed search model participants collaborate by sharing the information stored in their computers. Differently from the client-server model, P2P systems have the capability of increasing their performance as the number of users increases. To take advantage of this potential it is necessary to develop adaptive and collaborative mechanisms to exploit the semantics of users communities, the resources that they store and their search behavior. In order to address these issues, in this paper we present adaptive algorithms that learn to route queries to potentially useful nodes, reducing query propagation.

2 Background: Small World Topology and Semantic Communities

A good network logical topology is one that facilitates an effective performance and enables queries to reach the appropriate destiny in a few steps without overloading the system bandwidth [Tirado et al., 2010]. Moreover, it is desirable that the participants send their queries to other participants that are specialized in the query topic. Some results confirm this observation [Barbosa et al., 2004,Voulgaris et al., 2004]. This makes possible that a query be propagated quickly in the network through relevant nodes, and suggests that collaborative and distributed search can benefit from the context and the participants' community.

In order to evaluate the emergence of semantic communities in a P2P network we employ a methodology similar to the one applied in [Akavipat et al., 2006]. In particular, we adopt the concepts of "small world topology" and "clustering coefficient" [Watts and Strogatz, 1998] to study the structural properties of the emergent communication patterns.

2.1 Clustering Coefficient

The local clustering coefficient assesses the clustering in a single node's immediate network (i.e., the node and its neighbors) [Watts and Strogatz, 1998]. We consider undirected graphs $G = (V, E)$, in which V is the set of nodes and E is the set of edges. For a node v_i its neighborhood N_i is defined as the set of nodes v_j immediately connected to v_i, that is,

$$N_i =_{def} \{v_j \mid e_{ij} \in E, e_{ji} \in E\}$$

The local clustering coefficient is based on egos network density or local density. For each node v_i, this is measured as the fraction of the number of ties connecting v_i's neighbors over the total number of possible ties between v_i's neighbors.

Let k_i be the number of neighbors of a node v_i, that is, $|N_i|$. If a node has k_i neighbors then it could have at most $k_i(k_i-1)/2$ edges (if the neighborhood is fully connected).

Therefore, the local clustering coefficient for a node v_i can be formalized as follows:

$$C_i = \frac{2|e_{jk} \in E : v_j, v_k \in N_i|}{k_i(k_i-1)}.$$

In order to calculate the local clustering coefficient for the whole network, the individual fractions are averaged across all nodes [Watts and Strogatz, 1998]. Let n be the number of vertices in the network, that is $|E|$. Formally, the network average clustering coefficient can be defined as:

$$C_{average} = \frac{1}{n}\sum_{i=1}^{n} C_i.$$

A graph is considered *small-world* if its links are globally sparse (the network is far from being fully connected), its $C_{average}$ is higher than the average clustering coefficient associated with a random graph and the length of the path connecting two nodes is orders of magnitude smaller than the network size [Watts and Strogatz, 1998].

This metric represents the global knowledge of the network and was selected in this work in order to compare the ability of the proposed algorithms to understand the information associated with the nodes. When the amount of information about the nodes in the network is insufficient, $C_{average}$ is small. However, as this information grows, the value of $C_{average}$ will grow as well.

3 Algorithms

All the proposed algorithms share a common feature: each node has an internal table NT (Nodes' Topics) that contains the learned knowledge. Each entry maintains a topic and a set of nodes that are interested in this topic. The differences between the algorithms appear at the moment the table is updated and in the way a node is selected to send a query.

We designed eight context sensitive algorithms, adopting an incremental approach. Due to space limitations, in this paper we will only focus on the two algorithms that showed the best behavior. Despite showing small differences in the local behavior of each node, these two algorithms produced significant changes in the overall results. We will also present a brute-force algorithm as a baseline for comparative purposes.

3.1 Basic Algorithm

This algorithm does not have any intelligence, and therefore does not require the use of an NT table for each peer. The queries are routed in a brute-force search manner, as in Gnutella [Ripeanu, 2001]. Each time a node generates a query it

sends this message to all of the adjacent nodes. If a node that receives a query message can reply, it sends a reply message, otherwise, it forwards the query to its adjacent nodes until exhausting the initially defined number of query hops.

3.2 Adaptive Algorithm

In this algorithm at the moment of generating a new query message, the query-issuing node looks into its NT table for nodes associated with the topic of the query and sends the query message to all of them. In the case that the query-issuing node does not have an entry for this topic in its NT table, it sends the query message to all of the adjacent nodes, in the same way as the basic algorithm. The learning phase occurs with the reply message. When a node can reply a query it sends a reply message that follows the same path as the issued query. Each intermediate node in this path updates its NT table with the topic of the query that is being answered and the node that answered it. There is another component in this phase: updating messages. When a node learns something, after updating its NT table, it sends an update message with the information learnt –in the format (topic,node)– to all of its adjacent nodes and to all the nodes which are "known" (through its NT table) to be interested in the topic of the reply message. There is another situation in which a node must send update messages: when a query message arrives by broadcast and the node is interested in the topic of the query but cannot reply, it will send an update message to the node that originated the query. This behavior avoids excluding nodes that do not have many resources.

3.3 Selective Adaptive Algorithm

The only difference between this algorithm and the Adaptive Algorithm is that this version skips update messages to adjacent nodes and only sends this kind of messages to those nodes that are interested in the topic of the reply message that arrived by broadcast.

4 Simulations and Results

These algorithms were implemented in Java, the physical network was simulated with the OmNet++ framework [Pongor, 1993] and the logical network was visualized with JUNG (Java Universal Graph). As an input we used more than 40,000 scientific articles that were distributed among the nodes such that each node contained articles related with its interests.

To find the best algorithm for query routing ten simulations were performed with each one, storing the results of the first, third, fifth, seventh and tenth execution. In order to complete our analysis of the simulations we considered:

- The average clustering coefficient of the logical network.
- The number of queries that have been satisfied.

– The number of messages sent by each node taking into account update messages to analyze whether these kind of messages were congesting the network.
– The maximum number of hops needed to find an answer.

All the simulations were executed in a server with these characteristics:

– 32 processors (4 x 8 cores) Opteron.
– 50 GB RAM.
– Debian GNU/Linux 6.0 64 bits.
– kernel 3.8.3.
– Oracle JRE 1.7.0_21.

	Basic Algorithm					Adaptive Algorithm					Selective Adaptive Algorithm				
	1	**3**	**5**	**7**	**10**	**1**	**3**	**5**	**7**	**10**	**1**	**3**	**5**	**7**	**10**
Answered queries	$\frac{92}{150}$	$\frac{80}{150}$	$\frac{63}{150}$	$\frac{120}{150}$	$\frac{75}{150}$	$\frac{135}{150}$	$\frac{97}{150}$	$\frac{94}{150}$	$\frac{83}{150}$	$\frac{91}{150}$	$\frac{128}{150}$	$\frac{94}{150}$	$\frac{89}{150}$	$\frac{95}{150}$	$\frac{98}{150}$
Hops	30	29	29	28	30	25	2	2	2	2	29	2	2	2	2
Clustering coefficient	0.013	0.013	0.013	0.013	0.013	0.696	0.706	0.708	0.709	0.713	0.686	0.701	0.705	0.709	0.709
Sent messages (Millions)	0.996	0.998	0.998	0.998	0.995	1.964	1.462	1.367	1.345	1.139	1.898	1.447	1.409	1.391	1.186
Update messages (Millions)	–	–	–	–	–	0.214	0.141	0.129	0.128	0.127	0.201	0.126	0.124	0.119	0.114

Table 1. Performance comparison between algorithms

Table 1 presents the results that are considered more important for the comparison of the different algorithms. From this table we can conclude the following:

– The number of answered queries is higher in the basic algorithm and in the first execution of the other algorithms. This is because in these cases the queries are propagated through the whole network.
– On the other hand the maximum number of hops to find an answer decreases as the overall knowledge of the network increases.
– Related with the previous item, the average clustering coefficient increases in the latest executions. This is because the knowledge of the whole network is higher, so that the nodes can send a query directly to potentially useful nodes.
– Concerning the number of messages sent, we can see that this number decreases as the number of executions increases. This is because the nodes find their queries in fewer hops, so they need to propagate fewer messages.
– The update messages are a part of the sent messages. We can see that the Selective Adaptive Algorithm sends less update messages than the Adaptive Algorithm and the global knowledge is not modified.

It is important to distinguish the physical network from the logical one. We have established a physical network of 1000 randomly connected nodes that remains static through all the executions. Each node in this network is associated with one or more themes of interest. On the other hand, the logical network is the result of the evolution of the network's global knowledge. Through its graphical representation we can see the semantic communities that emerged from this incremental knowledge.

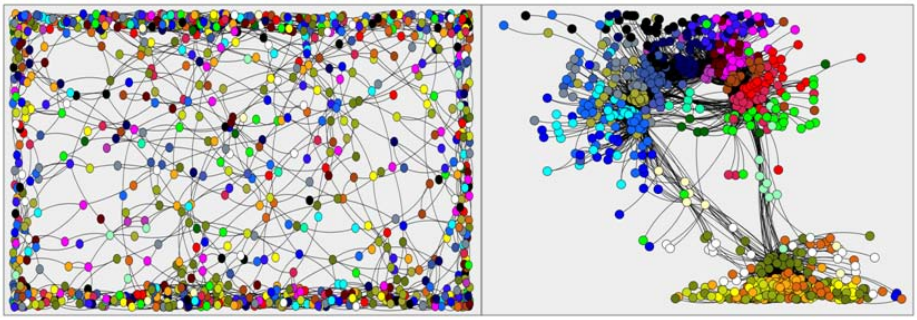


Fig. 1. Logical network obtained from a brute-force search algorithm (left). Logical network obtained from an intelligent algorithm (right).

On the left-hand side of figure 1 we can see a logical network obtained from the brute-force search algorithm. In this case the nodes only know their physical neighbors, disregarding the topics of interest associated with the rest of the pairs. This image shows us that this network does not reflect the existence of semantic communities. The logical network appearing on the right-hand side of figure 1 is the result of executing an intelligent algorithm. The colors are related with the topic in which each node is interested and allows us to appreciate the existence of semantic communities. It may be the case that in the physical network a pair of nodes is very far apart but the same pair is adjacent in the logical network. This is because the logical network reflects the semantic aspects of the nodes.

5 Related Work

A P2P system uses the computational power and the bandwidth of the participants instead of relying on a small number of servers [Balakrishnan et al., 2003]. Mechanisms for distributed content search in these systems offer solutions to some of the scalability and coverage problems commonly recognized in centralized systems. These limitations are particularly evident when attempting to design thematic portals, where small search engines attempt to offer solutions to specialized communities of users [Menczer et al., 2004, Pant et al., 2004].

Many investigations were done on how to structure the network for routing queries. A proposed solution is to create a two layer architecture: the upper is

the semantic layer that controls the super peers and the lower is the layer in charge of getting the relevant files [Athena Eftychiou, 2012]. Other approaches that use super peers were proposed in [Ismail et al., 2010], where decision trees are used in order to improve search performance for information retrieval in P2P network.

In the P2P scientific community there is an increasing interest in algorithms that dynamically modify the topology of the logical network, guided by mechanisms that allow the participants to learn about the thematic of the resources offered by other participants as well as their information needs [Wang, 2011, Yeferny and Arour, 2010]. This systems offers a way to relax the restrictions of centralized, planned and sequential control, resulting in decentralized and concurrent systems with collective behaviors [Watts and Strogatz, 1998].

There is a wide variety of search engines based on the P2P technology. For example the model proposed by the YouSearch [Bawa et al., 2003] project takes a centralized Napster-like design for query routing. At the same time each participant can find and index portions of the web. Other systems such as NeuroGrid [Joseph, 2002] attempt to send the query to potential nodes. Most of these systems use automatic learning techniques to adjust the metadata that describes the content of the nodes. Currently there exist some tools for decentralized search such as Faroo[1] and Yacy[2].

6 Conclusions and Future Work

The execution of the proposed algorithms in a simulation environment made it possible to obtain different statistics about their behavior such as response time and network congestion. With this statistical information we can conclude that the algorithms with better behavior are those that offer greater collaboration among peers (that is, when a node learns something, it should spread this knowledge across its community). Learning not only takes place to determine which node answers a query, but also when the node that generated the query is found to be semantically similar to the receptor node. In this case, learning occurs independently of whether the node replies or does not reply the query. These algorithms also showed that, after a number of executions, a logical network with a small-world topology and high average clustering coefficient emerges, reflecting the knowledge of the global network. The processing time that these algorithms require does not produce a significant overhead in the response time with the advantage that the processing time decreases as the available knowledge of the network increases.

Part of our future work will be focused on performing search based on semantic criteria, going beyond the currently existing syntactic search mechanisms. For example, if a query contains the term “house” an article that refers to a dwelling or an apartment could be also of interest for the user posing that query. This

[1] http://www.faroo.com.
[2] http://www.yacy.net.

kind of search by semantic similarity can reduce the precision but enables an increasing recall, reducing the number of ambiguities through context sensitivity.

A problem that arises in this scenario is what we could describe as "The Closed Communities Problem". In this setting, one or more nodes can be disconnected from their community or can form another community with the same topic without being related to each other. To solve this problem we plan to implement a curiosity mechanism that will prompt some participants to explore the network beyond their interest. Some results in this direction were already studied in [Maguitman et al., 2005,Lorenzetti and Maguitman, 2009]. Finally, we plan to run these algorithms in a real distributed environment where the participants could occasionally change their interests and generate queries dynamically. Research in this direction is currently underway.

7 Acknowledgements

We thank Maricel Luna and Mauro Ciarrocchi for their efforts in performing the initial analysis of the proposed algorithms. This research is funded by Projects PIP 112-200801-02798, PIP 112-200901-00863 (CONICET, Argentina), PGI 24/ZN10, PGI 24/N006 and PGI 24/N029 (SGCyT, UNS, Argentina).

References

Akavipat et al., 2006. Akavipat, R., Wu, L.-S., Menczer, F., and Maguitman, A. G. (2006). Emerging semantic communities in peer web search. In *Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, P2PIR '06, pages 1–8, New York, NY, USA. ACM.

Athena Eftychiou, 2012. Athena Eftychiou, Bogdan Vrusias, N. A. (2012). A dynamically semantic platform for efficient information retrieval in P2P networks. *International Journal of Grid and Utility Computing*, Volume 3:271 – 283.

Balakrishnan et al., 2003. Balakrishnan, H., Kaashoek, M. F., Karger, D., Morris, R., and Stoica, I. (2003). Looking up data in P2P systems. *Commun. ACM*, 46:43–48.

Barbosa et al., 2004. Barbosa, M. W., Costa, M. M., Almeida, J. M., and Almeida, V. A. F. (2004). Using locality of reference to improve performance of peer-to-peer applications. *SIGSOFT Softw. Eng. Notes*, 29:216–227.

Bawa et al., 2003. Bawa, M., Bayardo, R. J., Jr., and Rajagopalan, S. (2003). Make it fresh, make it quick - searching a network of personal webservers. In *Proc. 12th International World Wide Web Conference*, pages 577–586.

Ismail et al., 2010. Ismail, A., Quafafou, M., Nachouki, G., and Hajjar, M. (2010). A global knowledge for information retrieval in P2P networks. In *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*, pages 229–234.

Joseph, 2002. Joseph, S. (2002). Neurogrid: Semantically routing queries in peer-to-peer networks. In *Proc. Intl. Workshop on Peer-to-Peer Computing*, pages 202–214.

Lorenzetti and Maguitman, 2009. Lorenzetti, C. M. and Maguitman, A. G. (2009). A semi-supervised incremental algorithm to automatically formulate topical queries. *Information Sciences*, 179(12):1881–1892. Including Special Issue on Web Search.

Maguitman et al., 2005. Maguitman, A. G., Menczer, F., Roinestad, H., and Vespignani, A. (2005). Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 107–116, New York, NY, USA. ACM.

Menczer et al., 2004. Menczer, F., Pant, G., and Srinivasan, P. (2004). Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, 4(4):378–419.

Pant et al., 2004. Pant, G., Srinivasan, P., and Menczer, F. (2004). Crawling the Web. In Levene, M. and Poulovassilis, A., editors, *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. Springer-Verlag.

Pongor, 1993. Pongor, G. (1993). Omnet: Objective modular network testbed. In *Proceedings of the International Workshop on Modeling, Analysis, and Simulation On Computer and Telecommunication Systems*, MASCOTS '93, pages 323–326, San Diego, CA, USA. Society for Computer Simulation International.

Ripeanu, 2001. Ripeanu, M. (2001). Peer-to-peer architecture case study: Gnutella network. In *Peer-to-Peer Computing, 2001. Proceedings. First International Conference on*, pages 99–100.

Tirado et al., 2010. Tirado, J. M., Higuero, D., Isaila, F., Carretero, J., and Iamnitchi, A. (2010). Affinity P2P: A self-organizing content-based locality-aware collaborative peer-to-peer network. *Computer Networks*, 54(12):2056–2070.

Voulgaris et al., 2004. Voulgaris, S., Kermarrec, A., Massouli, L., and van Oteen, M. (2004). Exploiting semantic proximity in peer-to-peer content searching. In *Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems*, pages 238–243, Washington, DC, USA. IEEE Computer Society.

Wang, 2011. Wang, L. (2011). Sofa: An expert-driven, self-organization peer-to-peer semantic communities for network resource management. *Expert Syst. Appl.*, 38:94–105.

Watts and Strogatz, 1998. Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Yeferny and Arour, 2010. Yeferny, T. and Arour, K. (2010). LearningPeerSelection: A query routing approach for information retrieval in P2P systems. In *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*, pages 235–241.

Red Pulsante con Aprendizaje Hebbiano para Clasificación de Patrones Ralos

Iván Peralta[1*], José T. Molas[1], César E. Martínez[1,2], and Hugo L. Rufiner[1,2,3]

[1]Laboratorio de Cibernética, Facultad de Ingeniería,
Universidad Nacional de Entre Ríos,
CC 47 Suc. 3, E3100, Ruta 11, km. 10, Oro Verde, Entre Ríos, Argentina
[2]Centro de I+D en Señales, Sistemas e Inteligencia Computacional (SINC(i)),
Facultad de Ingeniería y Cs. Hídricas, Universidad Nacional del Litoral
[3]CONICET, Argentina
[*]bioivanperalta@gmail.com

Resumen En las últimas décadas se ha intentado desarrollar Redes Neuronales Artificiales más realístas que intenten imitar con mayor precisión el funcionamiento de sus contrapartes biológicas. Es así como nacieron las Redes Neuronales Pulsantes. Uno de los principales usos de estas redes es la clasificación de patrones. Sin embargo su aplicabilidad en el mundo real ha sido limitada debido a la falta de métodos de entrenamiento eficientes. En este trabajo se presenta un nuevo modelo de red pulsante pensado para clasificar patrones ralos. El mismo puede entrenarse mediante reglas de aprendizaje hebbiano no supervisado. Se describe su estructura, funcionamiento y el algoritmo propuesto para su entrenamiento. Además, se reportan resultados de prueba con patrones generados artificialmente y se discute la factibilidad de su implementación en un dispositivo lógico programable tipo FPGA.

Keywords: neurona pulsante, red neuronal pulsante, algoritmo de aprendizaje, aprendizaje no supervisado, patrones ralos.

1. Introducción

Las redes neuronales artificiales, inspiradas por las neuronas biológicas, están compuestas por unidades básicas denominadas neuronas. Estas se encuentran interconectadas mediante distintos pesos que determinan la intensidad con que interactúan dichas neuronas. La búsqueda de una analogía más cercana a la realidad biológica ha dado lugar en las últimas dos décadas a la aparición de las denominadas *Redes Neuronales Pulsantes* (SNN: *Spiking Neural Networks*). El uso de las SNN se encuentra en crecimiento debido a su habilidad para afrontar diferentes problemas en distintas áreas tales como clasificación de patrones, control maquinal, procesamiento de imágenes, etc. Estas redes reproducen más fielmente los sistemas neuronales biológicos tratando, por un lado, de imitar la transferencia de información entre neuronas a través de pulsos tal como se realizan en las sinapsis biológicas con los Potenciales de Acción, como así también, el procesamiento dinámico de las señales dentro de las neuronas.

Se han desarrollado innumerables trabajos que utilizan las SNN en diferentes aplicaciones. En [2] se describe un ejemplo de aplicación en ingeniería biomédica en donde se analizan tres algoritmos de entrenamiento de una SNN para la detección de epilepsia y convulsiones a través de la clasificación de patrones de EEG el cual es un problema de reconocimiento de patrones complicado, en [3] se presenta MuSpiNN, otro modelo de SNN, para afrontar el problema anterior. También se han utilizado en sistemas de control, en [12] se utiliza una SNN para controlar los movimientos de un robot para evitar obstáculos mediante señales ultrasónicas. Más recientemente en [1] se diseña una red capaz de memorizar secuencias de eventos y en [6] se analiza una SNN dinámica para el reconocimiento de patrones espacio/espectro temporales. A su vez se ha demostrado que este tipo de redes puede mapearse con mayor facilidad que las redes tradicionales dentro de dispositivos lógicos programables tipo FPGA [9,10].

Este trabajo fue motivado por la necesidad de desarrollar un clasificador eficiente para un tipo especial de patrones, originados a partir de *representaciones ralas* de señales de interés. Este tipo de representaciones o códigos surgen al analizar las señales mediante diccionarios discretos que utilizan una gran cantidad de átomos, pero que describen cada una de ellas en términos de una pequeña fracción de estos elementos [8]. Así, la codificación lograda posee la mayoría de sus coeficientes igualados a cero (o casi cero) [5]. Dado que este trabajo se enfoca en el entrenamiento de una red pulsante para el reconocimiento de patrones ralos, sin importar el método para obtener dicha representación, se utilizarán inicialmente patrones ralos artificiales de tipo binario, generados en forma aleatoria para cada clase.

El resto del trabajo se organiza como se detalla a continuación. La Sección 2 introduce la estructura de la SNN, los modelos neuronales y su funcionamiento interno. La Sección 3 describe la codificación rala y las reglas de entrenamiento de la SNN. La Sección 4 describe el método, las condiciones experimentales de prueba de la red, los resultados obtenidos y una discusión sobre la factibilidad de reproducir este modelo de SNN en un dispositivo lógico programable tipo FPGA. Finalmente, la Sección 5 resume las conclusiones y posibles líneas de trabajo futuros.

2. Estructura de la SNN

2.1. Conexiones y tipos de neuronas

La estructura de la SNN consiste de dos capas de neuronas: una Capa Detectora de entrada y una Capa Integradora de salida. La primera capa se conecta por un lado con el vector patrón de entrada ν, el cual es un vector en $\mathbb{B}^N$ donde $\mathbb{B}$ es el conjunto $\{0, 1\}$, y por otro lado con la capa integradora. La cantidad de neuronas integradoras $\mathbf{I}$ es igual a la cantidad de detectoras $\mathbf{D}$ y se tienen tantas neuronas integradoras como clases de patrones $\mathbf{C}$ se deseen clasificar, es decir, $\mathbf{I} = \mathbf{D} = \mathbf{C}$. En la Figura 1 se presenta un esquema de la SNN.

Entre cada coeficiente ν_n del vector patrón ν y cada neurona d de la capa detectora existe un peso de interconexión W_{nd} que pondera que tan importante

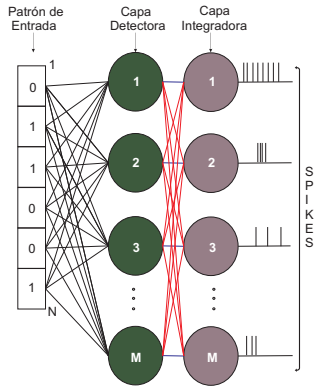


Figura 1. Diagrama estructural de la SNN, con N coeficientes para el vector patrón de entrada, M neuronas detectoras y M neuronas integradoras.

es la presencia de un "1" (uno) en dicho coeficiente para la clasificación del patrón por parte de la SNN. Cuanto mayor es el peso de interconexión, mayor importancia tendrá ese coeficiente para la clasificación del patrón dentro de alguna de las clases existentes. Si el coeficiente posee un valor "0" (cero) el mismo no estimula la neurona en la clasificación de ese patrón en particular. De la misma manera, existen pesos de interconexión entre la capa detectora y la capa integradora pero en este caso se diferencian dos tipos de conexiones: excitatorias –color azul– e inhibitorias –color rojo–. Cada neurona detectora excita a su correspondiente integradora a través de un peso positivo (W_{di} con $i = d$) e inhibe al resto mediante una conexión con pesos negativos (W_{di} con $i \neq d$). La comunicación entre las capas, así como también el vector de entrada con la capa detectora, se produce mediante spikes.

Cada neurona que integra la capa detectora recibe N conexiones; una por cada coeficiente del vector de entrada ν. Una neurona d de esta capa posee N registros internos correspondientes a los coeficientes v_n los cuales se denominan R_{nd}, a su vez posee un registro S_d que almacena la suma de los registros anteriores. Inicialmente todos estos registros se encuentran con valor cero.

El funcionamiento de la unidad esta temporizado en forma discreta con el tiempo k. Si en el instante k_0 se le presenta a la neurona detectora un vector patrón $\nu^{(0)}$, cada registro interno R_{nd} asociado a un coeficiente $\nu_n = 1$ se incrementará en el tiempo con una recta de pendiente igual al peso de interconexión entre el registro y el coeficiente. Este incremento en el registro se hará hasta llegar a un número preestablecido de iteraciones de subida k_f, luego cada registro regresará a cero en la iteración siguiente para esperar el próximo patrón de entrada $\nu^{(1)}$, a su vez antes de la presentación del próximo vector patrón se deja una iteración extra de espera. Por lo tanto, se presentará un vector patrón a la SNN cada $k_f + 2$ iteraciones y durante el período de tiempo entre una pre-

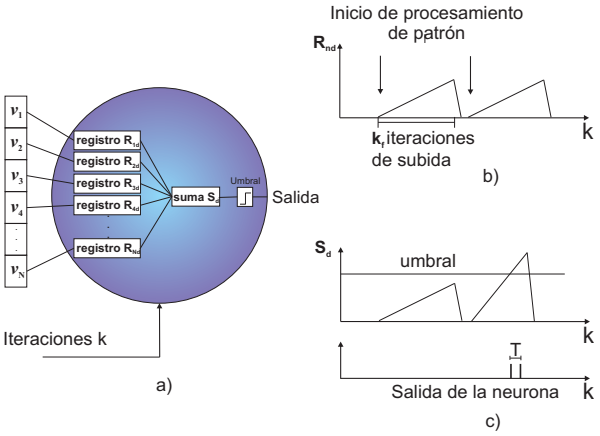


Figura 2. Diagrama estructural y funcional de la neurona detectora. a) representación de la estructura interna de la neurona. b) evolución de los registros una vez presentado un patrón con un $\nu = 1$ en el coeficiente correspondiente a ese registro. c) registro suma junto con el umbral, spikes a la salida y período refractario

sentación de un patrón y la siguiente, la neurona procesa dicha entrada para emitir pulsos ("spikes") en su salida si el registro S_d supera un cierto umbral. Una vez que la neurona emitió un pulso, existe un período refractario T en el que no se emiten pulsos por más que S_d se encuentre por encima del umbral. En la Figura 2 se resume la estructura y funcionamiento de la neurona detectora.

2.2. Neurona integradora

El modelo de neurona integradora propuesta para este trabajo posee una estructura muy similar a la neurona detectora, pero su funcionamiento interno es diferente. En la Figura 3 se describe el funcionamiento para la activación de varias de sus entradas. Cuando llega un pulso de alguna de las entradas, su registro R_{di} correspondiente se incrementará en el tiempo k con una recta de pendiente igual al peso de interconexión W_{di} entre estas neuronas, si $i = d$ el peso es positivo (conexión excitatoria) y si $i \neq d$ el peso es negativo (conexión inhibitoria). A diferencia del funcionamiento de la neurona detectora, este incremento en el registro se realiza sólo en una iteración, luego cada registro se acercará a cero pero con una pendiente W'_{di} inferior a W_{di} hasta recibir otro pulso en su entrada para volver a incrementarse. El registro S_i almacena la suma de todos los registros de esa neurona y si dicho registro supera un cierto umbral se emite un pulso en su salida. Al igual que en la neurona detectora, una vez que ésta emitió un pulso, existe un período refractario T en el que no se emiten pulsos por más que S_i se encuentre por encima del umbral. El valor del período refractario es el mismo para todas las neuronas integradoras pero es diferente del que poseen las neuronas detectoras.

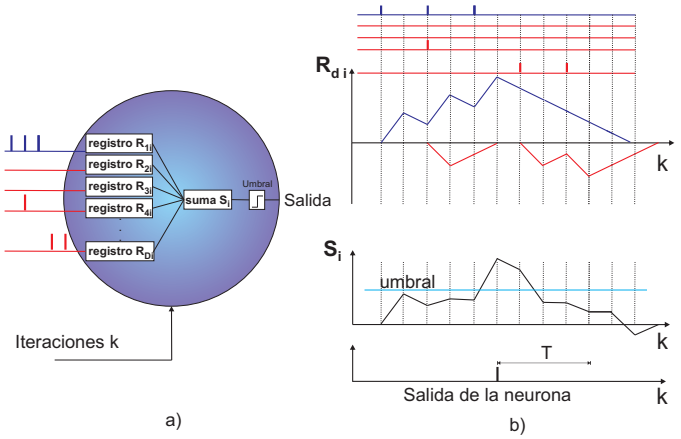


Figura 3. Diagrama Estructural y funcional de la neurona integradora. a) representación de la estructura interna de la neurona. b) integración de pulsos.

3. Algoritmo de aprendizaje

El método de aprendizaje de este trabajo se basa en la idea de Hebb [4]. El entrenamiento se realiza sobre las conexiones entre la capa detectora y los patrones de entrada, es decir, se determina el valor de los pesos W_{nd} descriptos en la Sección 2.1. Si se tiene una cantidad CE de vectores de cada clase para armar los patrones de entrenamiento y se denomina M_d a la matriz binaria de CE columnas y N filas formada por los patrones de entrenamiento de la clase d (recordar que $D = C$) en donde cada patrón forma un vector columna de la matriz. Si se denomina $Md(n, k)$ al n-ésimo coeficiente de la k-ésima columna de la matriz, entonces se puede describir la obtención de los pesos como:

$$W_{nd} = \begin{cases} \sum_{k=1}^{CE} M_d(n,k) & \text{si } \sum_{k=1}^{CE} M_d(n,k) > 0 \\ -\alpha \max_n \left\{ \sum_{k=1}^{CE} M_d(n,d) \right\} & \text{si } \sum_{k=1}^{CE} M_d(n,k) = 0 \end{cases} \quad (1)$$

donde α es una constante entera positiva.

La interpretación de esta regla es sencilla: cuanto más se active un coeficiente en los patrones de entrenamiento, mayor peso tendrá su conexión correspondiente. Si un coeficiente nunca se activó en el entrenamiento, entonces se le asigna un peso igual al negativo del máximo peso para esa clase escalado en un valor α. Esto hace que durante la evaluación cuando se presente un patrón cuyos coeficientes con valor 1 coincidan en su mayoría con las activaciones más frecuentes ocurridas durante el entrenamiento para alguna de las clases, el registro suma

de la neurona detectora correspondiente a esa clase supere el umbral y el patrón sea detectado.

3.1. Función de las neuronas integradoras

La neurona detectora emitirá mayor cantidad de pulsos si el patrón que está analizando corresponde con la clase de dicha neurona, dado que alcanzará su umbral más rápidamente que en caso contrario. La función de la neurona integradora consiste en integrar los pulsos provenientes de la detectora, esta integración maximizará el registro S_i en el caso de que el patrón pertenezca a la misma clase de este par de neuronas.

A menudo en la clasificación de patrones es necesario tener en cuenta la historia de los patrones que se van presentando a la red a lo largo del tiempo, un ejemplo de este tipo de clasificación es cuando la SNN está destinada a detectar patrones pertenecientes a clases que simulan distintos fonemas los cuales pueden abarcar varios patrones o solo uno [11]. Otra función de la neurona integradora consiste en recordar la clase de los patrones anteriores al analizado actualmente. Cuanto más leve sea la pendiente de decaimiento W'_{di}, mayor influencia tendrá el actual patrón en la detección de los siguientes pero si W'_{di} es muy pequeña, los registros S_i tardarán varias iteraciones en regresar a 0 y producirán pulsos que interferirán en la detección de patrones de otras clases. En el caso de la Figura 3, luego de los tres primeros pulsos de la entrada excitatoria vienen dos pulsos de una de las entradas inhibitorias, lo que supone que primeramente se presentó un patrón perteneciente a la misma clase de esa neurona y luego otro de una clase diferente. Si el valor de W'_{di} fuese más pequeño, entonces el valor del registro excitatorio permanecería elevado y podría generar pulsos a la salida e interferir en la detección del patrón de la segunda clase. Los pesos W_{di} negativos aumentan la especificidad de la detección (envío de pulsos inhibitorios hacia neuronas vecinas).

3.2. Determinación de los umbrales

Un punto importante es la determinación de los umbrales de los dos tipos de neuronas, sus valores no pueden ser muy elevados dado que no se tendría sensibilidad en la detección de los patrones. Por otro lado, si son muy pequeños se perdería especificidad entre las distintas clases de patrones. Cada neurona detectora posee un umbral proporcional al máximo valor alcanzado por sus registros R_{nd} durante todo el entrenamiento:

$$umbral_d = \beta k_f \max_n \{W_{nd}\} \tag{2}$$

donde $\beta < 1$ es la constante de proporcionalidad.

Cuanto mayor sean las iteraciones de subida k_f de una neurona detectora, mayor cantidad de pulsos excitarán la neurona integradora correspondiente y mayor el valor de sus registros internos. Por lo tanto, se fijó el umbral para estas neuronas como una proporción de k_f y es el mismo para todas las neuronas:

$$umbral_d = \gamma k_f \tag{3}$$

donde $\gamma > 1$ es la constante de proporcionalidad.

4. Experimentos y resultados

Como se mencionó anteriormente, en este trabajo no se aborda el problema de la obtención de los patrones ralos a partir de señales reales, sino que estos son generados en forma aleatoria y utilizados para entrenar y evaluar la SNN. Esto significa que se necesitan dos conjuntos de patrones: entrenamiento y evaluación. A su vez tendremos C clases o grupos de patrones. Los patrones utilizados se generaron artificialmente a partir del siguiente algoritmo:

1. Inicialmente se genera un conjunto de vectores binarios en $\mathbb{B}^N$ donde $\mathbb{B}$ es el conjunto $\{0, 1\}$ y N es un entero. Este inicio se hace favoreciendo la aparición de ceros respecto a la aparición de unos en el vector, esto lo hace un patrón ralo.
2. Se quitan aquellos vectores que posean todos sus elementos nulos.
3. Se aplica el algoritmo de *K-Means* para agrupar los vectores en C clases distintas. Este método permite agrupar vectores en distintas clases buscando el grado de parecido entre dichos vectores [7].
4. Si la cantidad de patrones que obtuvo la clase con menos elementos, es inferior a la cantidad de los patrones de entrenamiento y evaluación necesarios: se vuelve al paso 1 con mayor cantidad de vectores iniciales.
5. Se toman CE vectores de cada clase para armar los patrones de entrenamiento y CT vectores de cada clase para armar los patrones evaluación.

Se efectuaron 30 realizaciones del mismo experimento para obtener un resultado promedio de la tasa de reconocimiento global. En la Figura 4 se muestra un conjunto de patrones de 60 coeficientes utilizados para una de las realizaciones.

Con el fin de evaluar el desempeño de la capa integradora, en cada experimento se introducen los patrones para prueba con distintas modalidades. Por modalidad se entiende la cantidad de patrones de una misma clase que se introducen en forma consecutiva, siendo variados entre 1 y 5. En el primer caso los patrones son ubicados en forma alternada de manera que no pueda entrar un patrón de la misma clase que el patrón que ingresó anteriormente, para este caso el efecto de integración no será realizado. En el último caso, la introducción de los patrones se hace con menor mezcla de las clases, es decir, primero se introducen 5 patrones de una clase, luego se introducen los 5 patrones de la otra clase y así hasta terminar de introducir la totalidad de los patrones de prueba, esperando una integración máxima. Dado que se están utilizando patrones aleatorios, no se intentó un ajuste preciso de los parámetros de la SNN. La determinación se hizo mediante experimentación y no se utilizaron técnicas para la optimización de dichos parámetros. En la Tabla 1 se detallan los parámetros utilizados.

En la Figura 5 se muestran las tasas de reconocimiento para cada realización y la tasa de reconocimiento promedio para todo el experimento, en función de

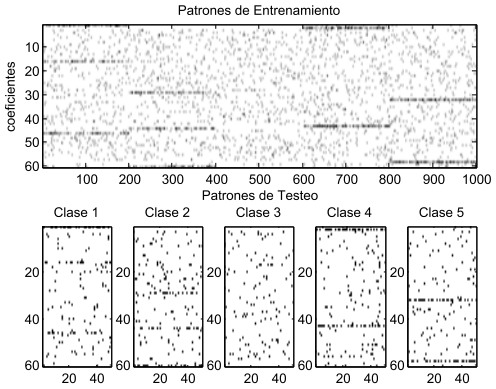


Figura 4. Patrones utilizados en una de las realizaciones del entrenamiento y prueba de la red. Arriba: 200 patrones de entrenamiento para cada clase. Abajo: 50 patrones de prueba para cada clase.

Tabla 1. Parámetros utilizados para la prueba de la SNN.

Referencia	Valor	Descripción
C	5	Cantidad de clases de patrones
D	5	Cantidad de neuronas de la capa detectora
I	5	Cantidad de neuronas de la capa integradora
k_f	8	Iteraciones de subida de la neurona detectora
T detectora	1	Periodo refractario de la neurona detectora
T integradora	4	Periodo refractario de la neurona integradora
W_{di}	16 (i == d)	Peso de interconexión entre capa detectora e integradora
	-13(i <> d)	
W'_{di}	5	Pendiente de descenso de los registros de la n. integradora
CE	200	Cantidad de patrones de entrenamiento por clase
CT	50	Cantidad de patrones de testeo por clase
α	4	Determinación de pesos neurona detectora
β	0,33	Determinación de umbral de neurona detectora
γ	1,5	Determinación de umbral de neurona integradora

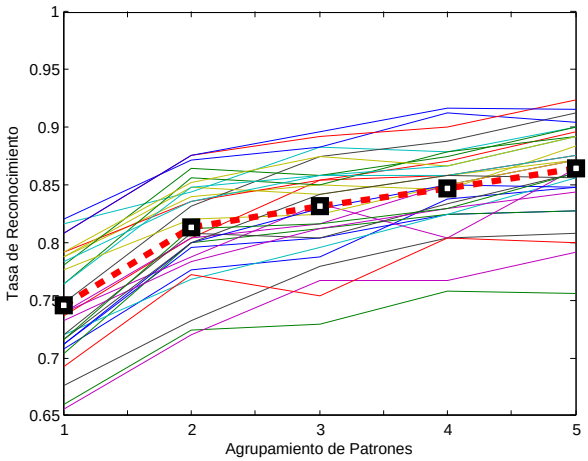


Figura 5. Tasas de reconocimiento en función de las 5 modalidades. En línea llena delgada: tasa de reconocimiento global de patrones para cada realización. En línea de trazos: tasa de reconocimiento promedio.

las 5 modalidades. Se aprecia una tasa de reconocimiento mínima superior al 65 % y máxima cercana al 83 % para la modalidad 1; mientras que la tasa de reconocimiento global promedio de todas las realizaciones se encuentra cercana al 75 %. A medida que aumenta la modalidad, las tasas de reconocimiento aumentan debido al funcionamiento de la capa integradora.

La utilización de pesos de interconexión entre capas de tipo enteros, las curvas de variación de tipo rectas así como la estructura basada en registros de esta SNN y la mayoría de los parámetros enteros, vuelven este diseño muy factible para mapearse en dispositivos electrónicos lógicos programables tipo FPGA con mínimos cambios. En [10] se ha desarrollado una SNN que posee las características mencionadas en una FPGA.

5. Conclusiones

En este trabajo se ha presentado el diseño e implementación de una Red Neuronal Pulsante, junto a un algoritmo de entrenamiento que permiten el reconocimiento de patrones ralos.

El desempeño obtenido es satisfactorio, sobre todo cuando se utilizan las capacidades de integración de estas redes mediante la presentación de patrones en forma consecutiva. Los resultados son bastante alentadores para la aplicación la cual fue motivación del presente trabajo.

Como trabajos futuros se puede mencionar que, con mínimos cambios, es posible trasladar la implementación lograda a un dispositivo portátil tipo FPGA.

Agradecimientos

Los autores desean agradecer a la *Agencia Nacional de Promoción Científica y Tecnológica* (bajo proyecto PAE 37122), la *Universidad Nacional de Entre Ríos* (PID NOVEL 6121), la *Universidad Nacional de Litoral* (PACT 2011 #58, CAI+D 2011 #58-511, #58-525), y al *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET).

Referencias

1. Borisyuk, R., Chik, D., Kazanovich, Y., Gomes, J.d.S.: Spiking neural network model for memorizing sequences with forward and backward recall. Biosystems (2013)
2. Ghosh-Dastidar, S., Adeli, H.: Improved spiking neural networks for EEG classification and epilepsy and seizure detection. Integrated Computer-Aided Engineering 14(3), 187–212 (2007)
3. Ghosh-Dastidar, S., Adeli, H.: A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection. Neural Networks 22(10), 1419–1431 (2009)
4. Hebb, D.O.: The organization of behavior: A neuropsychological approach. John Wiley & Sons (1949)
5. Hyvärinen, A.: Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation. Tech. rep., Helsinki University of Technology (1998)
6. Kasabov, N., Dhoble, K., Nuntalid, N., Indiveri, G.: Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition. Neural Networks (2012)
7. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, p. 14 (1967)
8. Olshausen, B., Field, D.: Emergence of simple cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
9. Pearson, M.J., Melhuish, C., Pipe, A.G., Nibouche, M., Gilhesphy, L., Gurney, K., Mitchinson, B.: Design and FPGA implementation of an embedded real-time biologically plausible spiking neural network processor. In: Field Programmable Logic and Applications, 2005. International Conference on. pp. 582–585 (2005)
10. Peralta, I., Molas, J.T., Martínez, C.E., Rufiner, H.L.: Implementación de una red neuronal pulsante parametrizable en FPGA. Anales de la XIV Reunión de Procesamiento de la Información y Control (2011)
11. Rufiner, H.L.: Análisis y modelado digital de la voz: técnicas recientes y aplicaciones. Ediciones UNL, Colección Ciencia y Técnica 284 (2009)
12. Wang, X., Hou, Z.G., Zou, A., Tan, M., Cheng, L.: A behavior controller based on spiking neural networks for mobile robots. Neurocomputing 71(4), 655–666 (2008)

A Cognitive Approach to Real-time Rescheduling using SOAR-RL

Juan Cruz Barsce[1], Jorge Palombarini[1,2], Ernesto Martinez[3]

[1] DEPARTAMENTO DE SISTEMAS (UTN) - Av. Universidad 450, X5900 HLR, Villa María, Argentina.
jbarsce@frvm.utn.edu.ar
[2] GISIQ(UTN) - Av. Universidad 450, X5900 HLR, Villa María, Argentina.
jpalombarini@frvm.utn.edu.ar
[3] INGAR(CONICET-UTN), Avellaneda 3657, S3002 GJC, Santa Fe, Argentina.
ecmarti@santafe-conicet.gob.ar

Abstract. Ensuring flexible and efficient manufacturing of customized products in an increasing dynamic and turbulent environment without sacrificing cost effectiveness, product quality and on-time delivery has become a key issue for most industrial enterprises. A promising approach to cope with this challenge is the integration of cognitive capabilities in systems and processes with the aim of expanding the knowledge base used to perform managerial and operational tasks. In this work, a novel approach to real-time rescheduling is proposed in order to achieve sustainable improvements in flexibility and adaptability of production systems through the integration of artificial cognitive capabilities, involving perception, reasoning/learning and planning skills. Moreover, an industrial example is discussed where the SOAR cognitive architecture capabilities are integrated in a software prototype, showing that the approach enables the rescheduling system to respond to events in an autonomic way, and to acquire experience through intensive simulation while performing repair tasks.

Keywords. Rescheduling, Cognitive Architecture, Manufacturing Systems, Reinforcement Learning, SOAR.

1 Introduction

In recent years, effective control of production systems is becoming increasingly difficult, because of growing requirements with regard to flexibility and productivity as well as a decreasing predictability of environmental conditions at the shop-floor. This trend has been accompanied by uncertainties in terms of an increasing number of products, product variants with specific configurations, large-scale fluctuations in demand and priority dispatching of orders. In order to face global competition, mass customization and small market niches, manufacturing systems must be primarily collaborative, flexible and responsive [1] without sacrificing cost effectiveness, product quality and on-time delivery in the presence of unforeseen events like equipment failures, quality tests demanding reprocessing operations, rush orders, delays in material inputs from previous operations and arrival of new orders [2]. In this context, reactive scheduling has becomes a key element of any real-time disruption management strategy, because the aforementioned conditions cause production schedules and

plans ineffective after a short time at the shop floor, whereas at the same time opportunities arise for improving shop-floor performance based on the situation encountered after a disruption [2].

Most modern approaches to tackle the rescheduling problem involve some kind of mathematical programming to exploit peculiarities of the specific problem structure, bearing in mind prioritizing schedule efficiency [3] or stability [4]. More recently, Gersmann and Hammer [5] have developed an improvement over an iterative schedule repair strategy using Support Vector Machines. Nevertheless, feature based representation is very inefficient for generalization to unseen states, the repairing logic is not clear to the end-user, and knowledge transfer to unseen scheduling domains is not feasible [6]. In this way, many researchers have identified the need to develop interactive rescheduling methodologies in order to achieve higher degrees of flexibility, adaptability and autonomy of manufacturing systems [1, 2, 7]. These approaches require the integration of human-like cognitive capabilities along with learning/reasoning skills and general intelligence in rescheduling components. Therefore, they can reason using substantial amounts of appropriately represented knowledge, learn from its experience, explain itself and be told what to do, be aware of its own capabilities and reflect on its own behavior, and respond robustly to surprise [8]. In this way, embodying continuously real-time information from the shop-floor environment, the manufacturing system and the individual product configuration through abstract sensors, rescheduling component can gain experience in order to act and decide in an autonomic way to counteract abrupt changes and unexpected situations via abstract actuators [9, 10].

In this work, a novel real-time rescheduling approach which resorts to capabilities of a general cognitive architecture and integrates symbolic representations of schedule states with (abstract) repair operators is presented. To learn a near-optimal policy for rescheduling using simulated schedule state transitions, an interactive repair-based strategy bearing in mind different goals and scenarios is proposed. To this aim, domain-specific knowledge for reactive scheduling is developed and integrated with SOAR cognitive architecture learning mechanisms like chunking and reinforcement learning via long term memories [11]. Finally, an industrial example is discussed showing that the approach enables the scheduling system to assess its operation range in an autonomic way, and to acquire experience through intensive simulation while performing repair tasks in production schedules.

2 Real-time Rescheduling in SOAR Cognitive Architecture

In this approach, knowledge about heuristics for repair-based scheduling to deal with unforeseen events and disturbances are generated and represented resorting to using a schedule state simulator connected with the SOAR cognitive architecture [12]. In the simulation environment, an instance of the schedule is interactively modified by the system using a sequence of repair operators suggested by SOAR, until a repair goal is achieved or the impossibility of repairing the schedule is accepted. SOAR's solves the problem of generating and encoding rescheduling knowledge using a general theory of computation, which is based on goals, problem spaces, states and operators, which will be explained later in detail. To implement the proposed approach, the cognitive architecture is connected with the rescheduling component via .NET wrappers. In-

formation about the actual schedule state comes in via the perception module which is related to an InputLink structure, and is held in the perceptual short-term memory. Symbolic first order schedule state structures in the form of id-attribute-value are extracted from InputLink, and added to SOAR's working memory. Working memory acts as a global short-term memory that cues the retrieval of rescheduling knowledge from Soar's symbolic long-term memories, as well as being the basis for initiating schedule repair actions. The three long-term symbolic memories are independent, and each one of them uses separate learning mechanisms. Procedural long-term memory is responsible for retrieving the knowledge that controls the processing; such knowledge is represented as *if-then* production rules, which match conditions against the contents of working memory, and perform their actions in parallel. Production rules can modify the working memory (and therefore the schedule state). To control rescheduling behavior, these rules generate preferences, which are used by the decision procedure to select a schedule repair operator. Operators are a key component in this approach, because they can be applied to cause persistent changes to the working memory. The latter has reserved areas that are monitored by other memories and processes, whereby changes in working memory can initiate retrievals from semantic and episodic memory, or initiate schedule repair actions through the abstract actuator in the environment. Semantic memory stores general first order structures that can be employed to solve new situations i.e. if in schedule state s1 the relations `precedes(task1,task2)` and `precedes(task2,task3)` which share the parameter object "task2" are verified, semantic memory can add the abstract relation `precedes(A,B), precedes(B,C)` to generalize such knowledge. On the other hand, episodic memory stores streams of experience in the form of *state-operator…state-operator* chains. Such knowledge can be used to predict behavior or environmental dynamics in similar situations or envision the schedule state outside the immediate perception using experience to predict outcomes of possible courses of actions when repairing a schedule. Moreover, this approach uses two specific learning mechanisms associated with SOAR's procedural memory, i.e. *chunking* and *reinforcement learning* [13], for learning new production rules as the schedule repair process is performed, and tuning the repair actions of rules that creates preferences for operator selection. Finally, repair operators suggested by the SOAR decision procedure affect the schedule state and are provided to the real-time rescheduling component via the OutputLink structure.

3 Schedule States, Repair Operators and Rescheduling Goals representation in SOAR

SOAR's working memory holds the current schedule state, and it is organized as a connected graph structure (a semantic net), rooted in a symbol that represents the state. The non-terminal nodes of the graph are called identifiers, the arcs are called attributes, and the values are the other nodes. The arcs that share the same node are called objects, so that an object consists of all the properties and relations of an identifier. A state is an object along with all other objects which are substructures of it, either directly or indirectly. In the upper section of Figure 1, an example of a state named `<s>` in SOAR's working memory is shown. In this figure, some details have

been omitted for the sake of clarity, so there are many substructures and arcs in the state which are not shown.

In the situation depicted in Figure 1, we can see that in the working memory the actual state is called <s>, and it has attributes like avgTard with value 2.5 h., initTardiness with value 28.5 h. and a totalWIP of 46.83, among others. Also, there exists a resource attribute, whose value is another identifier, so it links state <s> with another objects called <r1>, <r2> and <r3>. <r1> corresponds to resource of the type extruder, it can process products of type A and B, with an accumulated tardiness in the resource of 17 h., and it has three tasks assigned to it: <t1>, <t11> and <t3>. These tasks are in turn objects in such a way that for instance <t1> has proper attributes to describe it like Product Type, Previous Task, Next Task, Quantity, Duration, Start, Finish and Due Date among others.

Similarly to tasks, resources have an important attribute called Processing Rate that determines the total quantity of a given type of product that the resource can process in one unit of time. Finally, a state attribute called Calculations stores data that is relevant to calculate tasks tardiness, resource tardiness, schedule total tardiness, and derived values.

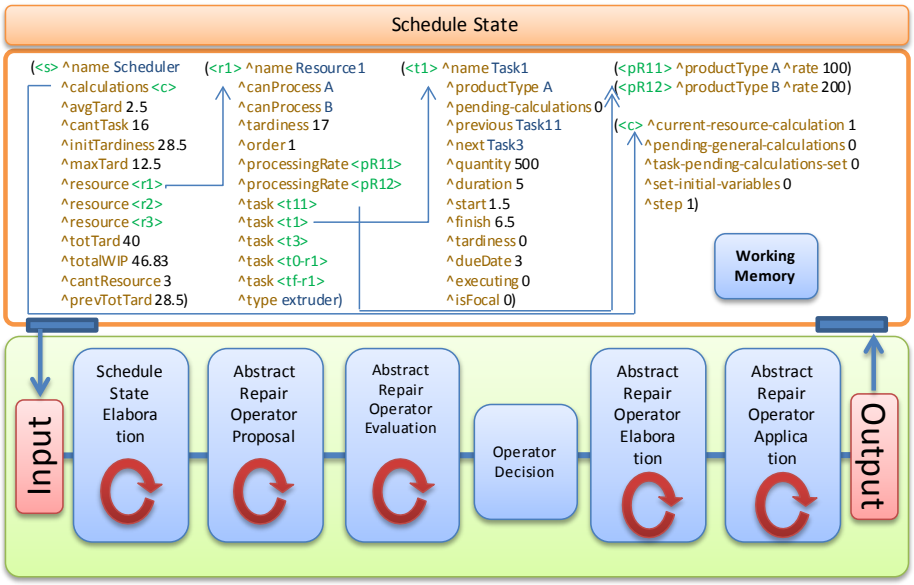


Fig. 1. Symbolic representation of schedule state (above) and repair operator elaboration, proposal and application cycle.

The concept of state is a key component to solve instances of rescheduling problems in SOAR, because SOAR programs are implicitly organized in terms of Problem Space Computational Model [14] so as to the conditions for proposing operators will restrict any operator to be considered only when it is relevant, and thus define the space of possible states that might be considered for achieving a repair goal. In this approach, the problem space is comprised of all possible schedules that can be generated when solving rescheduling tasks, and all repair operators that give rise to a schedule transition from one state to another. However, in a rescheduling task the

architecture does not explicitly generate all of feasible states exhaustively; instead, SOAR is in a specific schedule state at a given time (represented in working memory), attempting to select an operator to that will move it to a new, hopefully improved state. Such process continues recursively towards a goal state (i.e. schedule with a total tardiness that is minor than the initial tardiness). The lower section of Figure 1 shows the schedule repair process execution, which proceeds through a number of cycles. Each cycle is applied in phases; in the Input phase, new sensory data comes into the working memory and is copied as the actual schedule state. In the elaboration phase production rules fire (and retract) to interpret new data and generate derived facts (state elaboration). In the Proposal phase the architecture propose repair operators for the current schedule state using preferences, and then compare proposed operators (in evaluation phase). All matched production rules fire in parallel (and all retractions occur also in parallel), while matching and firing continue until there are no more additional complete matches or retractions of production rules (quiescence). Therefore, a decision procedure selects a new operator based on numeric preferences provided the reinforcement learning rules. Once a repair operator has been selected, its application phase starts and the operator application rules fire. The actions of these productions give rise to more matches or retract operations; just as during proposal, productions fire and retract in parallel until quiescence. Finally, output commands are sent to the real-time rescheduling component. The cycle continues until the halt action is issued from the Soar program (as the action of a production rule).

3.1 Design and Implementation of Repair Operators and Rescheduling Goals

As was explained in the previous section, repair operators are the way by which a schedule goes from one state to another until the rescheduling goal is reached. Hereby, deictic repair operators have been designed to move or swap a focal task with other tasks in the schedule (which may be assigned to other resource), so as to reach a goal state [10]. Each operator takes two arguments: the focal task, and an auxiliary task. Focal task is taken as the reparation anchorage point, and auxiliary task serves to specify the reparation action and evaluate its effectiveness. For example, if the proposed operator is `down-right-jump`, the idea is that the focal task must be moved to an alternative resource, and inserted after the auxiliary task. If so, the conditions of the `down-right-jump` operator proposal rule must assure that the auxiliary task has a programmed start time which is greater than the start time of the focal task before it has been moved. It is important to note that in the alternative resource may exist more than one task that meet the condition; in such case, the operator is proposed in parallel, parameterized with different auxiliary tasks. To exemplify the reasoning above, Figure 2 shows the `down-right-jump` application rule (left hand side at the left, and right hand side at the right). The left hand side of the rule in Figure 2 establishes the conditions that must be met by the schedule state so that the operator can be applied. In turn, the right hand side defines how the schedule state changes as a consequence of the repair operator application. All symbols enclosed in "<>" represent variables, and variables with the same name refer to the same object in both parts of the rule. Therefore, the rule in Figure 2 can be semantically expressed as: "if in the schedule, there exists a proposed operator named, `down-right-jump` which takes as argument a focal task named `<nametFocal>` and auxiliary task `<nametAux>`; also, there exists

a resource <r1> which has assigned tasks <tFocal>, <tPrevFocal> and <tPosFocal> which are different from each other, there exists a resource <r2> with a processing rate of <rater2> for the product type <prodType> which has assigned the different tasks <tAux> and <tPosAux>. The task <tFocal> is the focal task of the repair operator and its attributes take the values <quantity>, <duration> and <prodType>, the previous programmed task in the resource is <prevTFocal> and the next is <nexttFocal>. The task <tAux> is named <nametAux> and it is the auxiliary task in the repair operator and has as a previous programmed task named <prevtAux> and a next task <nexttAux>. Further, task <tPrevFocal> is named <nametPrevFocal>, task <tPosFocal> is named <nametPosFocal> and task <tPosAux> is named <nametPosAux>, then as a result of the operator application the schedule state change in the following manner: the new previous task of <tFocal> is <nametAux> and the next task is <nametPosAux>, the new next task of <tAux> is <nametFocal>, the new previous task of <tPosAux> is <nametFocal>, the new previous task of <tPosFocal> is <nametPrevFocal>, the new next task of <tPrevFocal> is <nametPosFocal>. Also, the new duration value of the focal task is (/ <quantity> <rater2>), the focal task is moved to resource <r2> and removed from <r1> and the value of pending-general-calculations is updated in 1.

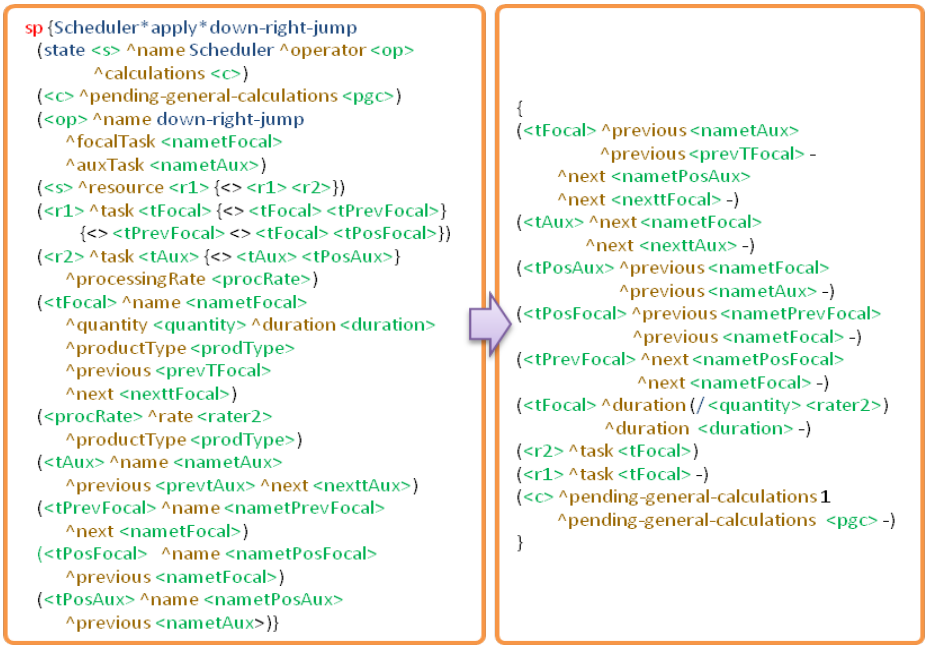


Fig. 2. Down-right-jump application rule.

As a result of the rule application the focal task has been moved to an alternative resource, its duration has been recalculated as from the processing rate of the alternative resource, and it has been inserted after the auxiliary task and the task originally programmed to start before of it. The pending-general-calculations value in 1 fires

new operators with the aim of recalculating task tardiness, resource tardiness, and other numeric values. An important matter is that once the repair operator and its arguments have been obtained, the rest of the variable values can be totally defined because they are related with each other, allowing an effective repair operator application. Another advantage of this approach relies on the use of variables in the body rule which act as universal quantifier, so the repair operator definition and application can match totally different schedule situations and production types only with the relational restrictions established in the left hand side of the rules.

Finally, after changes in the schedule state have been performed by the repair operator application rule, SOAR reinforcement learning rules are fired, so that the architecture can learn numeric preferences from the results of the particular repair operator application, which is carried on using a reward function and the SARSA(λ) algorithm [14]. The reward function is defined as the amount of tardiness reduced (positive reward) or increased (negative reward). Therefore, the SARSA(λ) algorithm updates the numeric preference of the operator using the well known formula in Eq. (1)

$$Q(s,ro)_{t+1} = Q(s,ro)_t + \alpha[r + \gamma Q(s',ro')_t - Q(s,ro)_t]e(s,ro)_t \quad (1)$$

where $Q(s,ro)$ is the value of applying the repair operator ro in the schedule state s, whereas α and γ are algorithm parameters, r is the reward value while $e(s,ro)$ is the eligibility trace value for repair operator ro in state s. Because of the problem space can be extremely large, and Q-values are stored in production rules which cannot be predefined, a reinforcement learning rule template [11] must be defined in order to generate updateable numeric preference rules that follow SOAR specifications for performing the learning procedure whenever visiting schedule states by means of available repair operators.

4 Industrial Case Study

An example problem proposed by Musier and Evans in [15] is considered to illustrate our approach for automated task rescheduling. It consists of a batch plant which is made up of three semi-continuous extruders that process customer orders for four classes of products (A, B, C and D). Each extruder can process one product at a time, and has its own characteristics. For example, the production rate for each type of product may vary from one extruder to another, and each extruder is not necessarily capable to process each type of product. Each task, in turn, has a due date, a required quantity (expressed in Kg.) and a product type.

Three applications have been used to implement and test the case study: *VisualSoar* v4.6.1, *SoarDebugger* 9.3.2 [12] and Visual Studio 2010 Ultimate running under Windows 7. Visual Studio 2010 was used to develop the real time scheduling component which allowed validating the results and read/write on the SOAR output/input link, respectively. *VisualSoar* environment was used to design and implement the definition of schedule state and operator proposal, elaboration and application knowledge.

Furthermore, *SoarDebugger* was used to run the aforementioned rules and train a rescheduling agent as well as to analyze the correctness of the operator pro-

posal/application rules. For the rescheduling problem space, there was a maximum of ten repair operators proposed for any state in each repair step and there are two classes of operators: *move* operators, which move the focal task into another position in the same resource or into an alternative one, and *swap* operators, which exchange the focal task with another task on different resources.

After each repair step, if the schedule has been repaired, the architecture is halted; otherwise, the agent propose/apply a new operator until the goal is achieved or an excessive amount of episodes has been made without finding the rescheduling solution. This situation may occur when the schedule to be repaired is very similar to the optimal schedule so further improvements are difficult to obtain Also, for each repair operator application, a reward is given to the agent based on how close the current schedule's tardiness is from the initial tardiness (i. e., how close is the repaired schedule to the goal state).

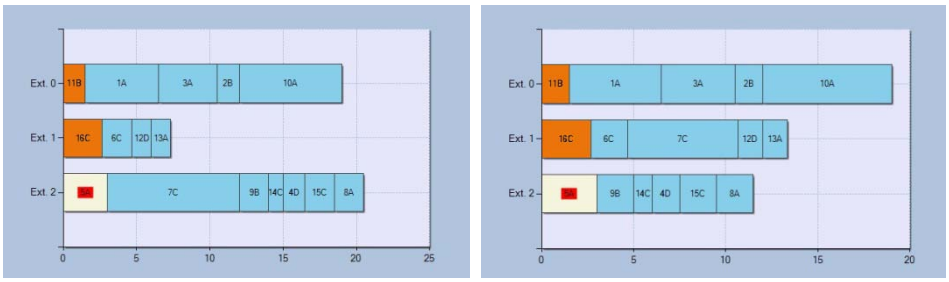
In this work, the disruptive event which has been considered is the arrival of a new order; for learning, the Sarsa (λ) algorithm was used with an ε-greedy police, eligibility traces, and parameters: $\gamma = 0.9$, $\varepsilon = 0.1$, $\lambda = 0.1$ and $\alpha = 0.1$. In the training phase of the rescheduling agent 20 simulated episodes were executed. In some episodes, the agent achieved the repair goal, which was stated as "*insert the arriving order without increasing the total tardiness present in the schedule before the order was inserted*". The intention behind such repair goal was on one hand, to insert the new order in an efficient way, and on the other, take advantage of the opportunity to improve schedule efficiency which may be caused by the very occurrence of a disruptive event [2] if it is possible. As the result of the training phase, 2520 reinforcement learning rules was generated dynamically. A representative example of one of such rules can be seen in Example 1 below (Some components of its left hand side have been omitted to facilitate reading and easy semantic interpretation).

```
Example 1. sp {rl*Scheduler*rl*rules*157
      (state <s1> ^totalWIP 46.83 ^taskNumber 16 ^maxTard 15
      ^avgTard 2.5 ^totTard 40 ^initTardiness 28.5 ^name
      Scheduler ^operator <o1> + ^focalTask <t1>)
      (<o1> ^auxTask Task10 ^focalTask Task5 ^name up-right-
      jump) --> (<s1> ^operator <o1> = -0.1498)}
```

The Example 1 rule was automatically instantiated by the SOAR cognitive architecture from an abstract learned rule, and is carried over the next schedule repair operations, so using these learned rules rescheduling decisions can be performed reactively in real time without extra deliberation. The rule in Example 1 reads as follows: if the Schedule state named <s1> has a Total Work in Process of 46.83, a Task Number of 16, a Maximum Tardiness of 15, an Average Tardiness of 2.5, a Total Tardiness of 40, an Initial Tardiness of 28.5, a Focal Task <t1> and the repair operator applied is up-right-jump, taken as auxiliary Task10 and Focal Task Task5, then the *Q*-value of that repair operator application is -0.1498. Evaluating such values for each operator, the agent can determine which one is the best in each situation, and act accordingly. Once the learning process has been performed, a new schedule was generated to test the learned repair policy.

To generate the initial schedule, 15 orders have been assigned to 3 resources as can be seen in Fig 3. Before the insertion of the focal task (highlighted in white), the schedule total tardiness was 28.5 h. After the insertion of the focal task the schedule

total tardiness has increased to 40 h. (see Fig. 4a). After six repair steps using the SOAR knowledge base, the initial schedule was successfully repaired, and the total tardiness was reduced to 18.5 h. (see Fig. 4b). Executing tasks at the time of insertion are highlighted in orange. The sequence of repair operators applied by SOAR was the following: the first operator was `up-right-jump` to Resource 1 (so Focal Task was inserted between Task 3 and Task 2), increasing the total tardiness to 44 h. That was followed by a `down-right-jump` to Resource 2 (behind Task 16), decreasing the total tardiness to 28.5 h. The third operator applied was an `up-right-jump` to Resource 1 (ahead of Task 10), increasing the total tardiness to 46.5 h. The next was a `down-right-jump` to Resource 3 (between Task 14 and Task 4). Next, SOAR applied an `up-left-jump` to Resource 2, (between Task 6 and Task 12). Finally, a `down-left-swap` to R3 was applied with Task 7, leaving the schedule with a total tardiness of 18.5 h.



(a). Initial Tardiness: 28.50 h. Tardiness After Insertion: 40.00 h.

(b).Final Tardiness after applying the repair policy: 18.50 h.

Fig. 3. Example of applying the optimal sequence of repair operators

As can be seen in the repair sequence, the rescheduling policy tries to obtain a balanced schedule using the focal task as the basis for swapping order positions in sequence of generated schedules, in order to take advantage of extruders with the best processing times. In this case, the rescheduling agent tries to relocate the Focal Task in an alternative resource so as to make a swap with Task 7 in order to move it to the second, sub utilized extruder. It is worth highlighting the small number of steps that are required for the scheduling agent to implement the learned policy in order to handle the insertion of an arriving order.

5 Concluding Remarks

A novel approach for simulation-based learning of a rule-based policy dealing with automated repair in real time of schedules using the SOAR cognitive architecture has been presented. The rescheduling policy allows generating a sequence of local repair operators to achieve alternative rescheduling goals which help coping with abnormal and unplanned events such as inserting an arriving order with minimum tardiness based on a symbolic first order representation of schedule states using abstract repair operators. The proposed approach efficiently represents and uses large bodies of symbolic knowledge, because it combines dynamically available knowledge for decision-

making in the form of production rules with learning mechanisms. Also, it compiles the rescheduling problem into production rules, so that over time, the schedule repair process is replaced by rule-driven decision making which can be used reactively in real-time in a straightforward way. In that sense, the use of abstract repair operators can match several non-predefined situations representing rescheduling tasks by means of problem spaces and schedule states using a relational symbolic abstraction which is not only efficient to profit from, but also potentially a very natural choice to mimic the human cognitive ability to deal with rescheduling problems, where relations between focal points and objects for defining repair strategies are typically used. Finally, by relying on an appropriate and well designed set of template rules, the approach enables the automatic generation through reinforcement learning and chunking of rescheduling heuristics that can be naturally understood by an end-user.

References

1. Zaeh, M., Reinhart, G., Ostgathe, M., Geiger, F., & Lau, C.: A Holistic Approach for the Cognitive Control of Production Systems. Advanced Engineering Informatics, 24, 300–307 (2010).
2. Aytug, H., Lawley, M., McKay, K., Mohan, S., Uzsoy, R.: Executing Production Schedules in the Face of Uncertainties: A Review and Some Future Directions. European Journal of Operational Research, 161, 86–110 (2005)
3. Adhitya, A., Srinivasan, R., Karimi, I. A.: Heuristic Rescheduling of Crude Oil Operations to Manage Abnormal Supply Chain Events. AIChE J. 53(2), 397-422 (2007)
4. Novas, J. M., Henning, G.: Reactive Scheduling Framework Based on Domain Knowledge and Constraint Programming. Computers and Chemical Engineering, 34, 2129–2148 (2010)
5. Gersmann, K., Hammer, B.: Improving Iterative Repair Strategies for Scheduling with the SVM. Neurocomputing, 63, 271–292 (2005)
6. Morales, E. F.: Relational State Abstraction for Reinforcement Learning. Proceedings of the Twenty-first Intl. Conference (ICML 2004), Banff, Alberta, Canada, July 4-8 (2004)
7. Henning, G.: Production Scheduling in the Process Industries: Current Trends, Emerging Challenges and Opportunities. Computer-Aided Chemical Engineering, 27, 23 (2009)
8. Brachman R. Systems That Know What They're Doing. IEEE Intelligent Systems, issue 6, 67-71 (2002)
9. Trentesaux, D.: Distributed Control of Production Systems. Engineering Applications of Artificial Intelligence, 22, 971–978 (2009).
10. Palombarini, J., Martínez, E.: SmartGantt – An Intelligent System for Real Time Rescheduling Based on Relational Reinforcement Learning. Expert Systems with Applications vol. 39, pp. 10251- 10268 (2012)
11. Nason, S., Laird, J. E.: Soar-RL: Integrating Reinforcement Learning with Soar. Cognitive Systems Research 6, 51–59 (2005)
12. Laird, J. E.: The Soar Cognitive Architecture, MIT Press, Boston (2012).
13. Nuxoll, A. M., Laird, J. E.: Enhancing Intelligent Agents with Episodic Memory. Cognitive Systems Research 17–18, 34–48 (2012)
14. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press (1998)
15. Musier, R., Evans, L.: An Approximate Method for the Production Scheduling of Industrial Batch Processes with Parallel Units. Computers and Chemical Engineering, 13, 229 (1989)

A Variant of Simulated Annealing to Solve Unrestricted Identical Parallel Machine Scheduling Problems

Claudia Gatica[1], Susana Esquivel[1] and Guillermo Leguizamón[1],

[1] LIDIC
Universidad Nacional de San Luis
Ejército de Los Andes 950 - Local 106
San Luis, Argentina
Telephone: (0266) 4420823
Fax: (0266) 4430224

{crgatica,esquivel,legui}@ unsl.edu.ar

Abstract. In this paper we propose a modification to the Simulated Annealing (SA) basic algorithm that includes an additional local search cycle after finishing every Metropolis cycle. The added search finishes when it improves the current solution or after a predefined number of tries. We applied the algorithm to minimize the Maximum Tardiness objective for the Unrestricted Parallel Identical Machines Scheduling Problem for which no benchmark have been found in the literature. In previous studies we found, by using Genetic Algorithms, solutions for some adapted instances corresponding to Weighted Tardiness problem taken from the OR-Library. The aim of this work is to find improved solutions (if possible) to be considered as the new benchmark values and make them available to the community interested in scheduling problems. Evidence of the improvement obtained with proposed approach is also provided.

Keywords: Unrestricted Parallel Identical Machines Scheduling Problem, Simulating Annealing. Maximum Tardiness.

1 Introduction

The schedule of activities is a decision process that has an important role in production and multiprocessor systems, manufacturing and information environments, and transportation distribution[17]. In particular, this paper considers the unrestricted identical parallel machine scheduling problem in which the maximum tardiness has to be minimized. Objectives such as the completion time of the last job to leave the system, known as Makespan (C_{max}), is one the more important objective function to be optimized, because it usually implies high resource utilization. In different systems of real world it is also usual stress minimization of the due-date based objectives as Maximum Tardiness (T_{max}) among others. Branch and Bound and other partial enumeration based methods, which guarantee exact solutions, are prohibitively time consuming even with only 20 jobs. The parallel machine environment has been

studied for several years due to its importance both academic and industrial. The scheduling literature provides a set of dispatching rules and heuristics. Different metaheuristics have been used to solve scheduling problems. For example, the population-based metaheuristics such as Evolutionary Algorithms and Ant Colony Optimization [2], [4], [11]. The trajectory-based heuristics have also been applied to solve these types of problems. In [13] VNS was used to solve the makespan in uniform parallel machine scheduling problem with release dates. In other related work [1] the authors applied an Iterated Local Search metaheuristic to solve the unrestricted parallel machine with unequal ready time problem. In [18] VNS with an efficient search mechanism, is proposed to solve the problem of maximum C_{max} in unrelated parallel machine scheduling. A comparative study [19] was conducted between SA and GRASP to solve the problem of maximum C_{max} in single machine scheduling, there SA outperforms GRASP. A hybrid approach is addressed in [5] which integrates features of Tabu Search (TS), SA, and VNS to solve a parallel machine problem with total tardiness objective. Another hybrid approach is presented in [14] where the authors combine TS with VNS in a way that the TS algorithm is embedded into VNS acting as a local search operator for parallel machine scheduling problem. In [6], [7], and [8] the authors face the same problem with an approach involving Evolutionary Algorithms with multirecombination and insertion of specific knowledge of the problem.

The rest of this paper is organized as follows. The next section presents the scheduling problem. After this, in section 3, the proposed algorithm is described. Section 4 explains the experimental design. In section 5 the results are shown and discussed. Finally, in section 6 we present our conclusions and outline our future work.

2 Scheduling Problem

The problem we are facing can be stated as follows: there are n jobs to be processed without interruption on some of the m identical machines belonging to the system; each machine can process not more than one job at a time, job j $(j = 1, 2, ..., n)$ is made available for the processing at time zero. It requires an uninterrupted positive processing time p_j on a machine and it has a due date d_j by which it should ideally be finished. For a given processing order of the jobs (schedule) the earliest completion time C_j and the maximum delay time $T_j = \{C_j - d_j , 0\}$ of the job j can readily be computed. The problem is to find a processing order of the jobs with minimum objective values. The objective to be minimized is:

$$\text{Maximum Tardiness: } T_{max} = max_j (T_j) \quad (1)$$

This problem is NP-hard when $2 \leq m \leq n$ [17].

3 The Proposed SA Algorithm

In a previous study [3] we work on the same problem but we address it with different local search metaheuristics: SA, VNS, Iterated Local Search (ILS) and Greedy Random Adaptive Search Procedure (GRASP). This comparative study showed that the best algorithm was SA although it was only able to improve benchmark values in ten instances (see Table 1). For reasons of space only the results obtained for m = 5 and n= 100 are showed. From the results obtained we assumed that the algorithm lacked of higher exploration capacity. With the main idea of overcoming these difficulties, we design a variant of the SA algorithm.

I	Bench	ILS	GRASP	VNS	SA
1	548	587	597	**547**	**542**
6	1594	1594	**1581**	**1572**	**1567**
11	2551	2577	2626	2552	**2539**
19	3703	3756	3784	3717	3718
21	5187	5193	5232	5197	**5177**
26	84	148	407	101	**70**
31	1134	1160	1366	1145	1135
36	2069	2128	2360	2091	**2061**
41	3651	**3631**	3821	**3621**	**3607**
46	4439	4475	4599	4443	4440
56	617	725	1104	655	**609**
61	1582	1779	2453	1705	**1580**
66	2360	2483	2870	2427	**2359**
71	3786	3924	4413	3862	3791
86	1194	1455	2281	1393	1194
91	2204	2427	2953	2412	2222
96	3185	3256	3780	3216	3187
111	1365	1846	3216	1781	1458
116	2222	2537	3055	2457	2266
121	2999	3407	3890	3286	3099

Table 1: Best values achieved by each metaheuristic

The pseudo-code of the proposed SA algorithm is given in Algorithm 1. The search processes of our algorithm is divided into two stages, based on the equilibrium condition as follows: SA starts with a high initial temperature (IT = 14256), it generates a random initial solution, and it initializes the counter to the equilibrium condition, which is achieved with the length of the Markov chain (LMC = 9716), which represents a constant number of search steps that are performed without updating the temperature (T). The justification for the initial value of temperature (IT), the length of the Markov chain (LMC) as the selection of operators (op1 and op2) is given in subsection 4.2 Then, depending on the condition of equilibrium, the search process is divided into two stages. In the first stage, the solutions are generated through the perturbation operator (op1 = scramble) (step: 7) and in the second stage, once the equilibrium condition is reached, and before updating the temperature (step: 16) it applies an extra exploration procedure called Explore (step: 15) which is described in Algorithm 2. Algorithms 1 and 2 show schematically the search process performed SA.

Algorithm 1 SA Algorithm including a call to an exploration procedure.

```
1: c = 0  {Used for the equilibrium condition}
2: s = s0 {Initial solution}
3: T = T0 {Starting temperature}
4: repeat
5:    repeat
6:          c  = c + 1
7:          Generate a solution s0 applying a
            perturbation operator (op1)
8:          ΔE = f(s0) – f(s)
9:          if ΔE ≤ 0 then
10:                 s = s0
11:         else
12:                 Accept s0 with a probability e^(-ΔE/T)
13:         end if
14:   until c == Markov-chain-length
15:   s0 = Explore(s)
16:   Update (T) {Geometric temperature update}
17:   c = 0
18:   ΔE = f(s0) – f(s)
19:   if ΔE ≤ 0 then
20:        s = s0
21:   else
22:        Accept s0 with a probability e^(-ΔE/T)
23:   end if
24: until Stopping Criteria
25: return s
```

Algorithm 2 Explore(s): the exploration procedure.

```
1: Input: s solution from SA, tries is the number of attempts
2: i = 1
3: while i ≤ tries do
4:  Generate a solution s0 applying a perturbation
    operator (op2)
5:  if f(s0) < f(s) then
6:   s = s0
7:   return s
8:  else
9:   i = i + 1
10: end if
11: end while
12: return s
```

The function Explore performs (i = 1, ... , *tries*) attempts to find a solution s_0 that improves s, as follows: generates a solution s_0 by applying a perturbation operator (*op2* = 4-*opt*). If f(s_0) < f(s), s is replaced by s_0 and Explore returns s, otherwise, another attempt is made by (steps : 4−6).

Following Algorithm 1, the acceptance criteria is applied (steps: 8 − 13 and 18 - 23). The search process ends when it reaches a maximum number of evaluations (step: 25).

In our implementation, the representation of the solutions is a permutation of integers in the range 1... *n,* which represent the job indexes.

The initial solution is a integer permutation randomly generated as follows: from 1 to n, for each index i generates a random number between i and n. This process checks that the solution is a valid representation, i.e. it is a permutation without repetition.

4 Experimental Design

4.1 Instances for the Unrestricted Parallel Identical Machines Scheduling Problem

Unlike other scheduling problems as Flow Shop or Job Shop, after an intensive search in the literature we could not find significant benchmarks for the problem we worked on. With the purpose of creating our own benchmarks, we extracted value pairs (p_j, d_j) based on selected data corresponding to Weighted Tardiness problem taken from the OR Library [10]. The values p_j and d_j correspond to the processing time and due date, respectively. These data were taken from problem sizes of 40 and 100 jobs. For each problem size, twenty instances were selected, each one with the same identification number although they were not the same problem, i.e., we had a problem numbered 1 with 40 jobs, another 1 with 100 jobs, and so on.

#I	*m*=2, *n*=40		*m*=5, *n*=40	
	DR	**MCMP-SE**	**DR**	**MCMP-SE**
1	235 (EDD)	**216**	284 (SLACK)	**230**
6	599 (SLACK)	**595**	652 (SLACK)	**606**
11	1060 (EDD)	**998**	1130 (SLACK)	**1016**
19	1628 (EDD)	**1624**	1700 (SLACK)	**1639**
21	1660 (SLACK)	**1634**	1720 (SLACK)	**1647**
26	55 (EDD)	**35**	100 (SLACK)	**61**
31	494 (EDD)	**474**	644 (SLACK)	**546**
36	869 (SLACK)	**852**	984 (SLACK)	**887**
41	1280 (EDD)	**1271**	1340 (EDD)	**1317**
46	1240 (EDD)	**1195**	1310 (SLACK)	**1235**
56	247 (SLACK)	**229**	318 (SLACK)	**252**
61	604 (EDD)	**604**	737 (SLACK)	**669**
66	1090 (SLACK)	**1071**	1240 (SLACK)	**1129**
71	1280 (EDD)	**1254**	1330 (SLACK)	**1272**
86	493 (SLACK)	**457**	589 (SLACK)	**508**
91	896 (EDD)	**874**	1040 (EDD)	**955**
96	1537 (EDD)	**1531**	1690 (SLACK)	**1607**
111	659 (EDD)	**621**	794 (SLACK)	**689**
116	650 (SLACK)	**627**	810 (SLACK)	**695**
121	1430 (EDD)	**1377**	1580 (SLACK)	**1469**

Table 2: Obtained values for 2 - 5 machines and 40 jobs

The numbers of the instances are not consecutive because each one was selected randomly from different groups. The tardiness factor is harder for those with the highest identification number.

These instances are available on request (email: crgatica@unsl.edu.ar). In a previous work [7], those value pairs were used as input for different dispatching rules (SPT: Shorted Processing Time first, EDD: Earliest Due Date first, SLACK: Least Slack, HODG Algorithm, and R&M: Rachamadugu and Morton Heuristic) provided by PARSIFAL [17], a Software Package provided by Morton and Pentico, and a Multi Crossover Multi Parent Genetic Algorithm (MCMP-SE) with insertion of knowledge [8]. The results obtained are showed in Table 1 (cases *m*=2, *n*=40 and *m*=5, *n*=40) and Table 2 (cases *m*=2, *n*=100 and *m*=5, *n*=100).

#I	*m*=2, *n*=100		*m*=5, *n*=100	
	DR	**MCMP-SE**	**DR**	**MCMP-SE**
1	562 (EDD)	**536**	590 (SLACK)	**548**
6	1550 (EDD)	**1544**	1680 (SLACK)	**1594**
11	2560 (EDD)	**2516**	2620 (SLACK)	**2551**
19	3690 (SLACK)	**3679**	3720 (SLACK)	**3703**
21	5150 (EDD)	**5143**	5240 (SLACK)	**5187**
26	60 (R&M)	**21**	168 (SLACK)	**84**
31	1110 (SLACK)	**1092**	1180 (SLACK)	**1134**
36	**2040** (SLACK)	2041	2120 (SLACK)	2069
41	3590 (EDD)	**3576**	3710 (SLACK)	**3651**
46	4420 (EDD)	**4396**	4580 (SLACK)	**4439**
56	582 (HODG)	**556**	670 (SLACK)	**617**
61	1560 (EDD)	**1549**	1630 (SLACK)	**1582**
66	2360 (EDD)	**2313**	2440 (SLACK)	**2360**
71	3780 (EDD)	**3741**	3820 (SLACK)	**3786**
86	1200 (EDD)	**1153**	1240 (SLACK)	**1194**
91	2180 (SLACK)	**2132**	2230 (EDD)	**2204**
96	3110 (SLACK)	**3093**	3250 (SLACK)	**3185**
111	5340 (WLPT)	**1325**	1420 (SLACK)	**1365**
116	2200 (EDD)	**2164**	2320 (SLACK)	**2222**
121	2940 (EDD)	**2934**	3060 (SLACK)	**2999**

Tabla 3: Obtained values for 2 - 5 machines and 100 jobs

In both Tables, #I indicates the instance identification and DR stands for Dispatching Rules. In the case of the dispatching rules, the displayed values correspond to the best obtained by the different rules used, whose names are enclosed in brackets. Bold values from both tables are considered as benchmarks in the present work.

4.2 Parameter Settings

In this subsection we describe the method used to determine the set of appropriate parameter values for our metaheuristic. There are different ways to do this, but can distinguish two main groups of techniques: one, when the sample used is formed with extreme values of the design space (*no space-filling*) or otherwise, when data values correspond to the interior of the design space (*space-filling*) [21]. The latter approach is the one we choose because it assumes that the interior of the design space can meet important characteristics of the true design model. For the generation of the samples we use the method Latin Hypercube Design (LHD) which generates random points within the design space. For the SA algorithm and Explore function their relevant parameters and corresponding application ranges were determined. They are indicated in Table 4. We use five different operators of disturbance or movement: n swaps (1), 2-opt (2), 4-opt (3), shift (4) and scramble (5). A detailed description of these operators can be found in [22]. Then LHD was employed using 20 design points which resulted in 20 different parameter configurations, this task was performed using the statistical tool R [20]. The resulting points sampling are shown in Table 5.

LMC=Length Markov chain	[1000, 10000]
CR=Cooling Rate	[0.5, 1.0]
IT=Initial Temperature	[10000, 100000]
OP1=Pertubation Operator of SA	[1, 5]
OP2=Pertubation Operator of Explore	[1, 5]

NT=Number of Tries	[10, 20]

Table 4: Parameter Ranges

Ultimately, we perform 20 experiments. Each experiment consisted of 50 runs of the algorithm SA, each run with 300,000 evaluations of the objective function for each of the 20 instances of 100 jobs and 5 machines. For the statistical study we use a software tool proposed by [12]. Such is called CONTROLTEST and automatically applies various statistical tests, one of which is the Friedman test [15] and other post-hoc procedures [16]. Resulting from the application on the median values of the runs of different configurations allowed us obtain the Average Ranking of Friedman Test and so, we were able to establish that the best performers were the c_4 and c_8 configurations (See Table 5, in column RF, such corresponds to Average Ranking of Friedman test) and also we can conclude that there are not statistical significant differences between them because the corresponding adjusted *p*-values did not give values less than 0.05, see Table 6. The only difference of the behaviour of SA with the specified parameter setup for *c4* and *c8* (and the reason of selection of *c8* configuration) was the lowest number of evaluations used by SA to achieve the best values. For reasons of space, the tables showing these results are not given here.

Conf.	LMC	CR	IT	OP_1	OP_2	NT	RF
c1	1287	0,79391	86906	4	3	17	14,025
c2	6455	0,50691	57118	2	1	11	5,425
c3	2809	0,58518	93290	2	4	15	3,275
c4	**8258**	**0,66540**	**84705**	**5**	**2**	**16**	**1,775**
c5	3358	0,54591	30000	2	5	18	3,575
c6	4554	0,81334	59801	3	2	14	8,15
c7	4681	0,56859	69300	4	4	16	13,525
c8	**9716**	**0,61812**	**14256**	**5**	**3**	**11**	**1,575**
c9	8745	0,95200	54194	2	4	12	19,825
C10	5806	0,97936	42010	4	2	10	17,775
C11	3721	0,70923	20184	3	3	20	8,025
C12	7727	0,87080	67559	1	3	14	16,175
C13	1894	0,89262	14825	3	2	17	7,95
C14	9199	0,68212	73597	4	3	12	12,95
C15	6071	0,75536	36234	2	4	15	5,525
C16	7246	0,93239	37356	3	2	18	18,625
C17	7903	0,83626	80125	2	5	13	17,1
C18	5348	0,64777	24275	4	4	19	12,65
C19	2336	0,74463	99708	1	1	13	11,875
C20	2615	0,92429	46197	3	2	19	10,2

Table 5: Parameter Configurations

config.	p-Bonf	p-Holm	p-Hoch	p-Homm
c4	1,05E+00	5,27E-01	5,27E-01	5,27E-01

Table 6: Adjusted *p*-values

4.3 Final Optimization Experiments

For each scenario, 50 runs were executed, each one with 600,000 objective function evaluations. In each experiment we calculate the following metrics:

1) **Best:** The best value found in each run.

2) **Median:** Is the median objective value obtained from the best found individuals throughout all runs.

3) **SD of Median:** The standard deviation of median objective value is the square root of its variance.

4) **Miter:** Is the mean of iterations where the best value was obtained.

5) **SD of Miter:** The standard deviation of mean of iterations in each run is the square root of its variance.

All the experiments reported in this work were run on a sub-cluster conformed by 1 CPUs of 64 bits, processor Intel Q9550 Quad Core 2.83GHz, with 4GB DDR3 1333Mz of memory, 500 Gb SATA and 2 TB SATA hard disks, Asus P5Q3 motherboard and 11 CPUs of 64 bits each with processor Intel Q9550 Quad Core 2.83GHz, 4GB DDR3 1333Mz memory, 160 Gb SATA hard disk and Asus P5Q3 motherboard.

5 Results and Discussion

For all cases studied, Table 7 synthesizes the best values of the objective function found by SA. In Table 7 entries marked in bold indicate that SA improved the benchmark value while entries in italic show that SA reached benchmark. For the case of 40 jobs and 2 machines, in almost all instances the benchmark values were achieved, except in instances 6, 26, and 116 where the algorithm was able to find smaller values.

	n=40				*n*=100			
	m=2		*m*=5		*m*=2		*m*=5	
#I	Bench	Best	Bench	Best	Bench	Best	Bench	Best
1	**216**	*216*	**230**	**229**	**536**	*536*	**548**	**539**
6	**595**	**594**	**606**	**604**	**1544**	*1544*	**1594**	**1569**
11	**998**	*998*	**1016**	*1016*	**2516**	*2516*	**2551**	**2544**
19	**1624**	*1624*	**1639**	*1639*	**3679**	*3679*	**3703**	3708
21	**1634**	*1634*	**1647**	*1647*	**5143**	*5143*	**5187**	**5177**
26	**35**	**27**	**61**	**55**	**21**	*21*	**84**	**70**
31	**474**	*474*	**546**	**542**	**1092**	*1092*	**1134**	**1125**
36	**852**	*852*	**887**	**885**	**2041**	**2037**	**2069**	**2061**
41	**1271**	*1271*	**1317**	**1313**	**3576**	*3576*	**3651**	**3607**
46	**1195**	*1195*	**1235**	**1227**	**4396**	*4396*	**4439**	*4439*
56	**229**	*229*	**252**	**244**	**556**	*556*	**617**	**606**
61	**604**	*604*	**669**	**651**	**1549**	*1549*	**1582**	**1580**
66	**1071**	*1071*	**1129**	**1128**	**2313**	*2313*	**2360**	**2355**
71	**1254**	*1254*	**1272**	**1266**	**3741**	*3741*	**3786**	3791
86	**457**	*457*	**508**	**507**	**1153**	*1153*	**1194**	*1194*
91	**874**	*874*	**955**	**947**	**2132**	*2132*	**2204**	**2199**
96	**1531**	*1531*	**1607**	**1597**	**3093**	*3093*	**3185**	3187
111	**621**	*621*	**689**	**665**	**1325**	1331	**1365**	1397
116	**627**	**619**	**695**	**661**	**2164**	*2164*	**2222**	2264
121	**1377**	*1377*	**1469**	*1469*	**2934**	2939	**2999**	3089

Table 7: *Bench* and *Best* values found

For the case of 40 jobs and 5 machines SA in four instances (11, 19, 21, 121) obtained the same value as the benchmark. In all other instances found better values. Furthermore, in the scenario of 100 jobs and 2 machines, SA obtains a value less than the benchmark in instance 36. In the case of instance 121, the proposed algorithm does not reach the benchmark value but by a little difference; in all the remaining instances reaches the benchmark values. In the last case analyzed, 100 jobs and 5 machines, SA improves the benchmark values in 12 instances (1, 6, 11, 21, 26, 31, 36, 41, 56, 61, 66 and 91). In two instances, 46 and 86 matches the benchmark. It reaches values close to benchmark in instances 19, 71, and 96; but the values obtained in the instances 111, 116, and 121 are further away from the known values. Previously observed behaviours allow us to assume that SA behaves fairly well for problems that

involve more machines because it improves or reaches the known values of the objective function. In the case of 2 machines, it reaches in most instances the benchmark values and also produces some improvements. Since the true optimal values are unknown, we may not conclude categorically if the number of machines makes the problem harder or if we do not improve the benchmark is because these are the true optimum.

6 Conclusion

The parallel machine environment has been studied for several years due to its importance both academic and industrial. Unlike other scheduling problems we could not find significant benchmarks for the problem of our interest, so in previous works we created our own instances, for 40 and 100 jobs, extracting data from the OR-Library corresponding to Weighted Tardiness and then we adapt them for the T_{max} problem. The main objective of our work was propose an improved version of SA with additional exploration capabilities in order to find new benchmark values (when possible) on the 20 instances analyzed in each case. This objective was achieved for several considered scenarios, the improved version of SA found new benchmark values. These results encourage us to continue with our research in two main directions: *a*) discuss alternatives regarding the combination of trajectory-based metaheuristics (e.g., SA with VNS or GRASP and also SA with population-based metaheuristics), and *b*) increase the quantity of instances to be considered, by adapting instances of the Weighted Tardiness Problems available in the OR-Library in order to obtain an extended set of instances for future research.

Acknowledgments. The authors would like to thank to the Universidad Nacional de San Luis for its continuous support.

References

1. C. Chen, *An Iterated Local Search for Unrelated Parallel Machines Problem with Unequal Ready Times*, Proceedings of the IEEE International Conference on Automation and Logistics Qingdao, China September 2008.
2. C. Mihil□, A. Mih□il□, *An Evolutionary Algorithm for Uniform Parallel Machines Scheduling*, Second UKSIM European Symposium on Computer Modelling and Simulation, 978-0-7695-3325-4/08, 2008 IEEE DOI 10.1109/EMS.2008.34, 2008.
3. C. Gatica and S. Esquivel and G. Leguizamón, *Comparative Study of Trajectory Metaheuristics for the Resolution of Scheduling Problem of Unrestricted Parallel Identical Machines*, XVIII Congreso Argentino de Ciencias de la Computación, 2012.
4. C. Gatica and S. Esquivel and G. Leguizamón, *An ACO approach for the Parallel Machines Scheduling Problem*, Inteligencia Artificial 46(2010), 84-95, doi: 10.4114/ia.v14i46.1550, 2010.
5. D. Anghinolfi and M. Paolucci, *Parallel machines total tardiness scheduling with a new hybrid metaheuristic approach*, Computer Operations Res.34:3471-3490, 2007.
6. E. Ferretti and S. Esquivel, *Knowledge Insertion: An Efficient Approach to Simple Genetic Algorithms for Unrestricted for Parallel Equal Machines Scheduling*. GECCO'05, 1587-1588, 2005, Washington DC, USA.

7. E. Ferretti and S. Esquivel, *An Efficient Approach of Simple and Multirecombinated Genetic Algorithms for Parallel Machine Scheduling*, IEEE Congress on Evolutionary Computation, 1340-1347, September 2005, Scotland, UK, IEEE Centre.
8. E. Ferretti and S. Esquivel, *A Comparison of Simple and Multirecombinated Evolutionary Algorithms with and without Problem Specific Knowledge Insertion, for Parallel Machines Scheduling*, International Transaction on Computer Science and Engineering, 2005, volume 3, number 1, 207-221.
9. E. G. Talbi, *Metaheuristics from design to implementation*, by John Wiley & Sons, Canada, 2009.
10. J. E. Beasley, *OR-Library: distributing test problems by electronic mail*, Journal of the Operational Research Society 41 (11), 1990, pp 1069-1072, as mentioned on. *http://people.brunel.ac.uk/[squigle]mastjjb/jeb/info.html*.
11. J. Arnaout and R. Musa and G. Rabadi, *Ant colony optimization algorithm to parallel machine scheduling problem with setups*, 4th IEEE Conference on Automation Science and Engineering Key Bridge Marriott, Washington DC, USA August 23-26, p:578-582, 2008.
12. J. Derrac, S. Garcıa, D. Molina, F. Herrera, *A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms*, Swarm and Evolutionary Computation, 2011.
13. K. Li and B. Cheng, *Variable neighbourhood search for uniform parallel machine makespan scheduling problem with release dates*, 2010 International Symposium on Computational Intelligence and Design.
14. M. Sevaux and K. Sörensen, *VNS/TS for a parallel machine scheduling problem*, MEC-VNS: 18th Mini Euro Conference on VNS, 2005.
15. M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of American Statistical Association 3 (1937) 674-701
16. M. Friedman, *A comparison of alternative test of significance for the problem of the m rankings*, Annals of Mathematical Statistics 11 (1940) 86-92.
17. M. Pinedo, *Scheduling: Theory, Algorithms and System*, Prentice Hall, 1995.
18. N. Piersman and W. van Dijk, *A local search heuristic for unrelated parallel machine scheduling with efficient neighbourhood search*, Mathematical and Computer Modelling, vol. 24, no. 9, pp. 11-19, 1996.
19. P. Sivasankaran and T. Sornakumar and R. Panneerselvam, *Design and Comparison of Simulated Annealing Algorithm and GRASP to Minimize Makespan in Single Machine Scheduling with Unrelated Parallel Machines*, Intelligent Information Management, 2010, 2, 406-416, doi:10.4236/iim.2010.27050 Published Online July 2010 (http://www.SciRP.org/journal/iim).
20. R Project, *The R Project for Statistical Computing*, http://www.rproject.org/.
21. T. Bartz-Beielstein, *Experimental Research in Evolutionary Computation*, The New Experimentalism, Springer, 2006.
22. T. Bäck, D. B. Fogel, and Z. Michalewicz, *Handbook of Evolutionary Computation*, Institute of Physics Publishing Bristol Philadelphia and Oxford University Press, New York, USA, 1997.

Algoritmo evolutivo para el problema de planificación en proyectos de desarrollo de software

Germán Dupuy - Natalia Stark - Carolina Salto

Facultad de Ingeniería - Universidad Nacional de La Pampa
Argentina
germandupuy24@gmail.com - {nstark, saltoc}@ing.unlpam.edu.ar}

Resumen La planificación de tareas y la asignación de recursos en proyectos de desarrollo de mediana a larga escala es un problema extremadamente complejo y es uno de los principales desafíos de la gestión del proyecto, debido a su complejidad. El objetivo es minimizar la duración y el costo del proyecto. En este trabajo proponemos un algoritmo genético (AG) tradicional usando codificación binaria para representar una solución al problema de planificación de proyectos software. En particular nos centramos en la elección del operador de cruce, junto con su probabilidad; proponemos comparar el cambio en el rendimiento del AG al utilizar operadores genéticos tradicionales respecto de otros más específicos para el problema. Los experimentos mostraron que utilizar una recombinación tradicional es capaz de aumentar el rendimiento del algoritmo, manteniendo en niveles aceptables la velocidad de convergencia.

Keywords: proyectos de planificación software, algoritmos genéticos, cruce, probabilidades

1. Introducción

En la actualidad, la gestión de los proyectos de desarrollo de software requiere de la programación temporal, la planificación y la monitorización de tareas, asignando recursos de manera eficiente para conseguir objetivos específicos, cumpliendo con un conjunto de restricciones. En líneas generales, la planificación de proyectos software [1] consiste en determinar quién debe hacer qué, cuándo, con qué recursos y en qué momento. Las tareas pueden ser muy diversas: desde mantener documentos hasta escribir programas. Por recursos entendemos personal con sus habilidades y el tiempo. Los objetivos que se persiguen es minimizar la duración y el costo. Además de cumplir con dichos objetivos se debe asegurar una calidad mínima del producto [2].

Los objetivos en conflicto y las restricciones hacen de éste un problema de optimización NP-hard [11], para el cual no existen soluciones algorítmicas polinomiales conocidas, razón por la cual, las aplicaciones de apoyo disponibles en el mercado hacen un seguimiento pasivo del proyecto. Dado que para el administrador del proyecto sería de gran utilidad disponer de una herramienta que le

facilite obtener planificaciones de forma automática que concilien los objetivos en conflicto a nivel de proyecto, muchos investigadores recurrieron a metaheurísticas, en particular algoritmos genéticos (AGs), para resolver el problema [1,2,3,9]. El trabajo de Alba y Chicano [1] es probablemente el trabajo más conocido y representa el estado del arte en la resolución del problema usando AGs.

En este trabajo nos centramos en la resolución del problema de planificación de proyecto de software (PSP) usando algoritmos genéticos [8]. Sin embargo, es importante resaltar, que en algoritmos genéticos en general la elección de un operador de cruce y su probabilidad de aplicación constituyen aspectos críticos en el diseño del algoritmo [10], debido a que el uso de parametrizaciones inadecuadas frecuentemente producen una reducción importante en el rendimiento del algoritmo [12]. Esto se debe en general a que los espacios de búsqueda resultantes presentan características indeseables, como muchos óptimos locales y múltiples regiones factibles desconectadas, que hacen que el proceso de optimización resulte considerablemente más difícil.

El algoritmo genético propuesto para resolver el problema utiliza cromosomas binarios para representar soluciones al PSP. Estas soluciones son matrices que codifican la dedicación de cada empleado para realizar una determinada tarea. Por simplicidad en la implementación, estas matrices se traducen en cadenas binarias que manipula el algoritmo genético. Usar esta traducción permite aplicar distintos tipos de operadores de recombinación: operadores de cruce tradicionales (un punto, dos puntos, uniforme, entre otros) que actuarían directamente sobre la cadena binaria que representa la matriz solución o bien aquellos diseñados para la recombinación de matrices. Por lo tanto, el objetivo que se persigue en este trabajo consiste en determinar cuál operador genético se adapta mejor en la resolución del problema en cuestión, como así también determinar cuál es la probabilidad de cruce más adecuada, otro aspecto en conflicto relacionado con el operador de cruce. Hemos considerado un conjunto amplio de instancias, variando los distintos parámetros del algoritmo, para evitar sesgar las conclusiones y evitar la posibilidad de "hand-tuning" del algoritmo para una instancia particular del problema.

Este trabajo está organizado como sigue. En la Sección 2 se define el Problema de Planificación de Proyectos. En la sección 3 se trata la aplicación del AG al problema en cuestión, luego en la sección 4 se describe la parametrización utilizada, seguidamente se muestran los resultados en la sección 5 y por último, en la sección 6 se presentan algunas conclusiones y trabajos futuros.

2. Definición del Problema

El problema de planificación de proyectos (PSP, Project Scheduling Problem) plantea la necesidad de un procedimiento para asignar recursos limitados a un conjunto de tareas en un cierto plazo. Se está frente a un proceso de toma de decisiones que tiene como meta la optimización de objetivos que eventualmente pueden estar en conflicto entre sí. Los recursos gestionados son personas con un

conjunto de habilidades, un salario y un grado máximo de dedicación al proyecto. Los objetivos son, normalmente, minimizar la duración y el costo del proyecto.

La formulación del problema es como sigue [4]. Se tiene un conjunto de E empleados y un conjunto de T tareas con ciertas habilidades requeridas, además de un grafo de precedencia de tareas (TPG). Cada empleado posee un conjunto de habilidades, recibe un salario mensual y tiene un grado máximo de dedicación al proyecto (cociente entre la cantidad de horas dedicadas al proyecto y la cantidad de horas de una jornada laboral completa). Cada tarea está caracterizada por un conjunto de habilidades asociadas necesarias para poder ser llevadas a cabo y un esfuerzo expresado en personas-mes. Las tareas se deben completar según el orden establecido por el grafo de precedencia de tareas, el cual es un grafo dirigido que indica qué tareas deben completarse antes de iniciar otra.

El objetivo del problema es, entonces, asignar empleados a tareas, minimizando tanto el costo como la duración del proyecto. Se debe considerar que a cada tarea la realiza al menos una persona, las habilidades de los empleados que realizan cierta tarea deben cubrir el conjunto de habilidades requeridas por esa tarea y, por último, la dedicación de un empleado al proyecto no debe exceder su dedicación máxima. Si alguna de estas restricciones no se cumplen, se considera que la solución al problema no es factible. Por lo tanto, una solución al problema se puede representar por medio de una matriz $X = (x_{ik})$ de tamaño $E \times T$ donde $x_{ik} \geq 0$. El elemento x_{ik} es el grado de dedicación del i-esimo empleado a la k-esima tarea.

La calidad de una solución depende de tres factores: la duración y el costo del proyecto y la factibilidad de la solución. La duración de cada tarea (tkdur) es el cociente entre el esfuerzo para realizar la tarea y la sumatoria de los grados de dedicación de cada empleado a esa tarea. Obtenida la duración de cada tarea y teniendo en cuenta el TPG, se calcula el tiempo de inicio y fin de cada una, lo que da lugar a conocer la duración del proyecto, denotada por *tiempo*. El costo (*costo*) es el producto entre el salario del empleado y el tiempo de dedicación al proyecto; este último es la suma de la dedicación a cada tarea del empleado multiplicada por la duración de la misma.

3. Algoritmo Genético para el PSP

En este trabajo se utiliza un AG tradicional con representación binaria y reemplazo generacional, a diferencia de las propuestas de Alba y Chicano [1] que utilizan un AG de estado estacionario y de Minku et. al [9] que trabajan con una (1+1)-AG. La solución al problema se representa por una matriz X, donde el elemento x_{ij} indica el grado de dedicación del i-ésimo empleado a la j-ésima tarea. Se considera que el grado de dedicación máxima de cada empleado es igual a 1, ya que se supone que ningún empleado trabaja horas extras, es decir $x_{ij} \in [0, 1]$. Como la dedicación de cada empleado x_{ij} es un valor real, el intervalo de dedicación se discretiza en ocho valores uniformemente distribuidos, tomando para ello tres bits [1]. Como técnica de implementación y por simplicidad, se adoptó que la matriz solución X se represente como una cadena binaria en una

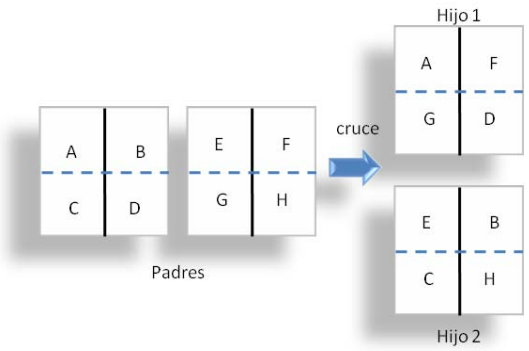


Figura 1. Funcionamiento del cruce de matrices

dimensión (x), es decir se realiza un procedimiento de aplanamiento de la matriz. En consecuencia, la dimensión de la cadena binaria x es igual a $E \times T \times 3$. La función de fitness que usa el AG para medir qué tan buena es una solución es la siguiente:

$$f(x) = \begin{cases} 1/q & \text{si la solución es factible} \\ 1/(q+p) & \text{en caso contrario} \end{cases} \quad (1)$$

donde $q = w_{costo} \times costo + w_{tiempo} \times tiempo$, $p = w_{penal} + w_{t\_inc} \times t\_inc + w_{h\_nocub} \times h\_nocub + w_{extras} \times extras$ y w_{costo}, w_{tiempo}, w_{penal}, w_{t_inc}, w_{h_nocub} y w_{extras} son parámetros predefinidos, *costo* y *tiempo* son el costo y el tiempo de finalización del proyecto asociados a la solución, *t_inc* es la cantidad de tareas que no tienen empleados asociados, *h_nocub* es la cantidad de habilidades requeridas no cubiertas para realizar todas las tareas del proyecto, y *extras* es el tiempo total de horas extras de todos los empleados durante la ejecución del proyecto. Como los dos términos q y p aparecen en el denominador y el objetivo es minimizarlos, entonces $f(x)$ se debe maximizar.

Un punto crucial en el diseño de un algoritmos genético es determinar un operador de cruce adecuado al problema. Dado que algunos son más disruptivos que otros, el tipo de cruce usado puede dirigir una búsqueda demasiado exhaustiva sobre una región o por el contrario hacer una gran exploración sobre el espacio de búsqueda. Como el AG propuesto en este trabajo maneja cromosomas representados como cadenas binarias, se propone utilizar el tradicional cruce de un punto para arreglos binarios (1X) introducido por Holland [7].

Por otra parte, como la solución al problema es una matriz, se decidió analizar el comportamiento del algoritmo utilizando un cruce diseñado especialmente para recombinar matrices, conocido como cruce 2D de un punto (2DMX) [13]. Luego de transformar la cadena binaria x en la matriz X, el cruce selecciona una fila y una columna (la misma en los dos padres) en forma aleatoria y luego intercambia los elementos en el cuadrante superior izquierdo y en el cuadrante

inferior derecho de ambos individuos. La Figura 1 muestra en forma gráfica el funcionamiento del cruce.

Analizando el grado de ruptura que podrían provocar ambos operadores, es de esperar que el operador 1X resulte con un menor grado de ruptura de la información (en este caso planificación) contenida en las soluciones padres ya que podría pasar filas completas (cada fila indica el grado de dedicación de cada empleado a las distintas tareas) de los padres al hijo. Por otra parte, como se puede apreciar en la Figura 1 la preservación de la información contenida en los padres es más compleja de lograr en el caso de que el AG utilice 2DMX.

La eficacia de un operador está muy relacionada con su probabilidad de aplicación (p_c). Por consiguiente, la elección de un valor adecuado de p_c afecta críticamente el comportamiento y rendimiento de los AGs. En particular la probabilidad de cruce controla la capacidad de los AGs para explotar un pico localizado con el objetivo de alcanzar un óptimo local. Cuanto más alta es la probabilidad de cruce, se da lugar a una rápida explotación. Pero, una p_c muy alta deberá disromper//perturbar los individuos en forma más rápida de lo que puedan ser explotados. Las configuraciones sugeridas para la probabilidad de cruce son $p_c = 0{,}6$ [5], $p_c = 0{,}95$[6] y $p_c \in [0{,}75, 0{,}95]$[12]. Estos valores fueron derivados de estudios empíricos sobre un determinado conjunto de problemas de prueba, y pueden ser inadecuado debido a que el valor de p_c es específico para el problema en estudio. Teniendo en cuenta estas consideraciones, en este trabajo se plantea analizar distintos valores de probablidades, desde bajos a altos, para determinar cuál de ellas es la más adecuada para los operadores estudiados, a fin de obtener un buen rendimiento del AG para resolver el problema de planificación de proyectos.

4. Parametrización

En esta sección se muestran los valores paramétricos utilizados para evaluar el rendimiento de las variantes del AG propuesto para resolver el problema de planificación de proyectos de software. Para llevar a cabo esta comparación se proponen las siguientes variantes de AG: (*i*) un AG con operador de cruce binario de un punto (AG_1X) y (*ii*) un AG con operador de cruce sobre matrices (AG_2DMX). Además, para evitar ajustar el algoritmo a una situación particular y permitir una comparación justa, por cada variante se propone el estudio de su comportamiento al utilizar distintas probabilidades de cruce, a saber: 0.3, 0.5, 0.75 y 0.9, es decir, de bajos valores de cruce a altos. De esta manera, resultan el estudio de ocho variantes algorítmicas (2 cruces por 4 probabilidades cada uno).

El tamaño de la población (μ) es de 64 individuos y es inicializada aleatoreamente. En cada generación se crean 64 individuos (λ). Como criterio de selección de padres se utiliza torneo binario y el reemplazo es ($\mu + \lambda$) utilizando selección proporcional por ruleta. El operador de mutación es el bit-flip con probabilidad igual a 0.005. El máximo número de evaluaciones está fijado en 20000.

Para realizar las experimentaciones y comparar los efectos de los operadores de cruce con las distintas probabilidades se tomaron las instancias propuestas

en [4]. Estas instancias representan una variedad de escenarios de la vida real. Los componentes de las instancias son: empleados, tareas, habilidades, y el grafo de precedencia de tareas (TPG). Las instancias empleadas representan proyectos que varían en el número de empleados: 5, 10 y 15, en el número de tareas: 10, 20 y 30, y en el número de habilidades de los empleados: 4 a 5, 6 a 7, 5 y 10 habilidades. Una de ellas podría ser por ejemplo un proyecto que cuenta con 10 empleados, 30 tareas, y 6 a 7 habilidades por empleado. De la combinación de habilidades, empleados y tareas se obtiene un total de 36 instancias. Los valores de los pesos de la función de fitness son [4]: $w_{costo} = 10^{-6}$, $w_{tiempo} = 10^{-1}$, $w_{penal} = 100$, $w_{t_inc} = 10$, $w_{h_nocub} = 10$ y $w_{extras} = 0{,}1$.

La experimentación realizada ha sido extensa: las ocho variantes algorítmicas en estudio se evaluaron con cada una de las instancias, en total se obtienen un total de 288 combinaciones (8 algoritmos $\times$ 36 instancias). Debido a la naturaleza estocástica de los algoritmos, cada una de ellas se ejecutan 30 veces para obtener una muestra confiable. En total se realizaron 30 $\times$ 288 ejecuciones haciendo un total de 8640 ejecuciones. Hemos realizado un análisis estadístico de los resultados a fin de obtener conclusiones significativas, en particular se aplicó el test no paramétrido Kruskal Wallis, considerando un nivel de significancia de $\alpha = 0{,}05$, para indicar un nivel de confidencia del 95 % en los resultados.

El algoritmo ha sido implementado con la librería MALLBA bajo C++. El entorno de ejecución consiste de máquinas con triple procesador AMD Phenom8450 a 2GHz, con 2 GB de RAM, bajo Linux, versión 2.6.27-4GB kernel.

5. Resultados

En esta sección se presentan y analizan los resultados obtenidos por las dos variantes de algoritmos genéticos presentados (AG_1X y AG_2DXM). En una primera etapa, se determina para cada variante cuál es la probabilidad de cruce que permite un mejor comportamiento del algoritmo. En una segunda fase, se procederá a comparar las mejores variantes a fin de determinar la configuración óptima del algoritmo.

Comenzamos el estudio con el examen del rendimiento de las variantes propuestas bajo distintas probabilidades de cruce. Al analizar los mejores resultados obtenidos para AG_1X, no se observa un claro patrón de comportamiento para determinar cuál probabilidad es la más adecuada para resolver el problema o grupo de instancias. Por tal motivo hemos recurrido a los estudios estadísticos para definir cuál probabilidad es la que brinda diferencias estadísticas significativas. El Cuadro 1 muestra los p-values del estudio estadístico realizado sobre AG_1X al comparar los valores medios obtenidos con las distintas probabilidades de cruce para cada una de las instancias del problema. Se puede observar que en 21 de las 36 instancias analizadas los valores son mayores a 0.05, el nivel de significancia, indicando que el comportamiento del AG no está influenciado por la probabilidad de cruce utilizada para resolver el problema PSP. En las restantes instancias (15 de 36) donde las diferencias entre los algoritmos es significante, se realizó un test de múltiple comparaciones que indica que AG_1X con $p_m = 0{,}3$

Cuadro 1. p-values del estudio paramétrico sobre las mejores soluciones de AG_1X discriminadas por instancia

empleados	habilidades	tareas		
		10	20	30
5	4-5	**0,265**	0,020	**0,051**
	6-7	**0,239**	0,018	**0,149**
	5	**0,590**	**0,153**	0,009
	10	0,000	0,001	**0,531**
10	4-5	0,000	**0,152**	**0,092**
	6-7	**0,130**	0,000	0,000
	5	**0,404**	0,008	0,044
	10	**0,360**	**0,327**	0,002
15	4-5	**0,331**	0,001	**0,081**
	6-7	0,015	0,016	**0,072**
	5	**0,627**	0,003	**0,875**
	10	**0,093**	**0,112**	**0,167**

marca las diferencias, resultando la variante algorítmica con el peor comportamiento. En cuanto al resto de las variantes planteadas, AG_1X con $p_m = 0{,}75$ y $p_m = 0{,}9$ no presentan rangos de valores medios con significancia estadística, pero para estas instancias en 8 de las 14 instancias 0.9 presenta valores medios más altos (lo que indica mejor calidad de resultados en promedio).

En cuanto a los resultados obtenidos con AG_2DXM, se observa una situación similar a la anterior referente a determinar en forma empírica la probabilidad que contribuyó a obtener los mejores resultados. A diferencia del análisis estadístico previo, en esta oportunidad los p-values son marcadamente menores al nivel de significancia 0.05, indicando que las distintas probabilidades arrojan resultados diferentes en todas las instancias analizadas (debido a esta homogeneidad de resultados es que no se ha incluido el cuadro correspondiente). El test de múltiples comparaciones sugiere que en comparación el rendimiento del algoritmo fue significantemente mejorado al usar una probabilidad de 0.9 en 27 de las 36 instancias. Nuevamente, la probabilidad que marca las diferencias es 0.3.

Concluimos, a partir de los resultados, que AG_1X y AG_2DXM obtienen sus mejores rendimientos cuando la probabilidad de cruce está configurada a 0,9. Esto sugiere que el algoritmo necesita altos valores de probabilidades para lograr obtener un buen rendimiento. Esta configuración de ambas variantes son las que se usarán para llevar a cabo las siguientes comparaciones.

El siguiente paso consiste en determinar cuál combinación de cruce y probabilidad es la que genera un mejor rendimiento del AG para resolver el problema de PSP. La Figura 2 muestra los valores de fitness de cada algoritmo para cada instancia. Como se puede observar AG_1X obtiene los mejores valores de fitness en la mayoría de las instancias (valores de fitness más altos). El test no paramétrico indica que en 30 de las 36 instancias hay diferencias significativas entre las dos variantes, dando soporte a las observaciones anteriores. De esta manera, el usar un cruce de un punto juega un rol importante para mejorar el rendimiento del algoritmo.

En función de la cantidad de generaciones promedio que cada variante necesita para alcanzar sus mejores soluciones (ver Figura 3), AG_2DXM presenta mayor velocidad de convergencia en 21 de las 36 instancias. Pero el estudio es-

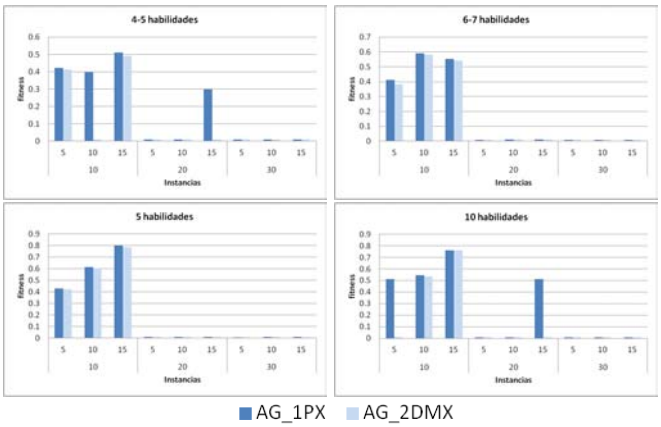


Figura 2. Mejores valores de fitness alcanzados por cada algoritmo, discriminados por instancia.

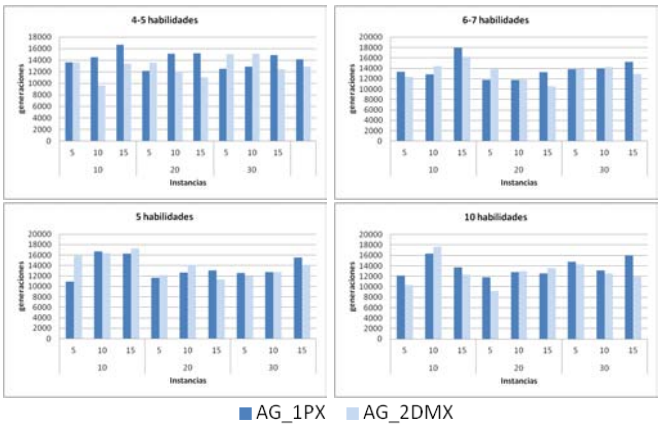


Figura 3. Evaluaciones promedio necesarias para obtener las mejores soluciones por cada algoritmo, descriminadas por instancia.

tadístico realizado indica que las diferencias son significativas sólo en 9 de las 36 instancias. Esto sugiere la superioridad de AG_1X sobre AG_2DXM para alcanzar mejores soluciones con una velocidad de convergencia similar.

Un aspecto importante a analizar es el porcentaje de éxito de cada variante, determinada por el porcentaje de soluciones obtenidas que son factibles, es decir soluciones que cumplen con todas las restricciones impuestas en la resolución del problema (ver Sección 2). El Cuadro 2 muestra el porcentaje de las mejores soluciones de cada algoritmo que son factibles (cero indica que no se encontraron soluciones factible). El porcentaje de factibilidad está totalmente vinculado con la cantidad de tareas que se deben realizar en cada proyecto: cuanto mayor es

Cuadro 2. Porcentaje de soluciones factibles de cada algoritmo AG_1X y AG_2DXM

		10 tareas		**20 tareas**		**30 tareas**	
empleados	habilidades	AG_1X	AG_2DXM	AG_1X	AG_2DXM	AG_1X	AG_2DXM
5	4-5	**100**	**100**	0	0	0	0
	6-7	**29**	21	0	0	0	0
	5	71	**93**	0	0	0	0
	10	0	0	0	0	0	0
10	4-5	**64**	0	0	0	0	0
	6-7	**100**	**100**	**7**	0	0	0
	5	86	**100**	0	0	0	0
	10	**100**	**100**	0	0	0	0
15	4-5	**100**	36	**21**	0	0	0
	6-7	**100**	**100**	0	0	0	0
	5	**100**	**100**	0	0	0	0
	10	14	**64**	**100**	36	0	0

la cantidad de tareas menor es el porcentaje de factibilidad obtenido por ambos algoritmos. En el caso de proyectos con 10 tareas ambos algoritmos encuentran soluciones factibles para todas las intancias, en 9 de las 12 instancias el porcentaje de AG_1X está por encima del 60 %, caso similar se observa para AG_2DXM (8 sobre 12 instancias). También vale resaltar que en 7 instancias alguno de los dos algoritmos obtienen en todas las ejecuciones soluciones factibles (porcentaje de éxito del 100 %). Para proyectos con 20 tareas AG_1X obtiene mayor porcentaje de soluciones factibles que AG_2DXM. Comparando nuestros resultados con los porcentajes de factibilidad obtenidos en [4] y en [9], observamos que son similares. Una observación importante es que en las instancias que representan proyectos con 30 tareas, los porcentajes de éxito mostrados en la literatura son cercanos a 0, no presentando importantes diferencias con nuestra propuesta.

6. Conclusiones

Este trabajo presenta una opción evolutiva para determinar la asignación de tareas a empleados, con el objetivo de minimizar el tiempo de duración del proyecto. La opción evolutiva es un algoritmo genético tradicional con representación binaria. El estudio consistió en analizar el comportamiento de este algoritmo bajo dos tipos de cruce: el tradicional operador de un punto para representaciones binarias y un operador de cruce a nivel matricial ya que las soluciones al problema son matrices que representan la dedicación de cada empleado a las tareas. También se analizaron cuatro valores distintos de probabilidades de cruce ya que es un parámetro influyente en el rendimiento de un algoritmo. Los resultados muestran que para este problema es conveniente utilizar probabilidades de cruce altas. La variante algorítmica utilizando el operador tradicional obtuvo el mejor rendimiento, obteniendo buena calidad de patrones de planificación. Los porcentajes de factibilidad de las soluciones obtenidas son altos para problemas de planificación no tan complejos, estos valores son similares a los obtenidos por otras variantes evolutivas de la literatura para resolver el problema en cuestión.

Como trabajo futuro proponemos el desarrollo de un operador genético más específico del problema o un algoritmo de búsqueda local que permita mejorar

los niveles de factibilidad de las soluciones obtenidas, para que tengan sentido para aquellos administradores que llevan a cabo un proyecto de desarrollo de software. En este sentido, también se pondrá atención en el mecanismo de manejo de restricciones, tal como está planteado se usa un mecanismo de penalización pero se podrían considerar mecanismos más avanzados. Además del estudio experimental para proponer otros pesos para la función de fitness.

Reconocimientos

Los autores agradecen el apoyo de la UNLPam y ANPCYT (proyecto 09F-049). Germán Dupuy agradece al CIN por la beca de EVC 2012.

Referencias

1. E. Alba and J. F. Chicano. Software project management with GAs. *Information Sciences*, 177:2380–2401, 2007.
2. C. K. Chang, M. J. Christensen, and T. Zhang. Genetic algorithms for project management. *Annals of Software Engineering*, 11:107–139, 2001.
3. C. K. Chang, H. Jiang, Y. Di, D. Zhu, and Y. Ge. Time-line based model for software project scheduling with genetic algorithms. *Information and Software Technology*, 50(11):1142–1154, 2008.
4. J.F. Chicano. *Metaheurísticas e Ingeniería del Software.* PhD thesis, University of Málaga, 2007.
5. K. DeJong. *An analysis of the behavior of a class of genetic adaptive systems.* PhD thesis, University of Michigan, Ann Arbor, MI, 1975.
6. J. J. Greffenstette. Optimization of control parameters for genetic algorithms. *IEEE Transaction on System Man and Cybernetic*, 16(1):122–128, 1986.
7. J.H. Holland. *Adaptation in natural and artificial systems.* The University of Michigan Press, 1975.
8. M. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer, third revised edition, 1996.
9. L.L. Minku, D. Sudholt, and X. Yao. Evolutionary algorithms for the project scheduling problem: runtime analysis and improved design. In *Proceedings of International Conference on Genetic and Evolutionary Computation Conference*, GECCO '12, pages 1221–1228, 2012.
10. G. Ochoa, I. Harvey, and H. Buxton. On recombination and optimal mutation rates. In *Proceedings of Genetic and Evolutionary Computation Conference*, pages 488–495. Morgan Kaufmann, 1999.
11. L. Ozdamar and G. Ulusoy. A survey on the resource-constrained project scheduling problem. *IIE Transactions*, 27:574–586, 1995.
12. J.D. Schaffer, R.A. Caruana, L.J. Eshelman, and R. Das. A study of control parameters affecting online performance of genetic algorithms for function optimization. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 51–60, 1989.
13. B.M. Wall. *A genetic algorithm for resource-constrained scheduling.* PhD thesis, MIT, 1996.

Algoritmos Evolutivos Multirecombinativos Híbridos Aplicados al Problema de Vehículos con Capacidad Limitada

Viviana Mercado[1], Andrea Villagra[1] , Daniel Pandolfi[1] , Guillermo Leguizamón [2]

[1] Laboratorio de Tecnologías Emergentes LabTEm, Departamento de Ciencias Exactas y Naturales, Universidad Nacional de la Patagonia Austral Unidad Académica Caleta Olivia,
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional, Departamento de Informática, Universidad Nacional de San Luis
[1]{vmercado, avillagra, dpandolfi }@uaco.unpa.edu.ar, [2]legui@unsl.edu.ar

Resumen. El objetivo perseguido en este campo es fundamentalmente el desarrollo de nuevos métodos capaces de resolver problemas complejos con el menor esfuerzo computacional posible, mejorando así a los algoritmos existentes. Las metaheurísticas son métodos que integran procedimientos de mejora local y estrategias de alto nivel para realizar una búsqueda robusta en el espacio-problema. El problema de ruteo de vehículos es un problema de optimización combinatoria de gran importancia en diferentes entornos logísticos debido a su dificultad (NP-duros). Se han propuesto varias soluciones a este problema haciendo uso de heurísticas y metaheurísticas. En este trabajo proponemos dos algoritmos para resolver el problema de ruteo de vehículos con capacidad limitada, utilizando como base un Algoritmo Evolutivo multirecombinativo conocido como MCMP-SRI (Stud and Random Inmigrants), combinado con operadores de mutación basados en conceptos computación cuántica. Detalles de los algoritmos y los resultados de los experimentos muestran un promisorio comportamiento para resolver el problema.

Keywords: Metaheurísticas, Algoritmos Evolutivos multirecombinativos, Hibridación, Problema de Ruteo de Vehículos.

1 Introducción y Trabajos Relacionados

El problema de ruteo de vehículos (Vehicle Routing Problem o las siglas en inglés VRP) consiste en generar rutas de reparto dado una cantidad de clientes por atender, un conjunto de vehículos de reparto y un punto de origen, permitiendo minimizar ciertos factores que ayuden a la empresa a obtener beneficios [4] y [12].
Algunas de las metaheurísticas más comúnmente utilizadas en el problema de ruteo de vehículo y sus variantes son por ejemplo los Algoritmos Genéticos (AGs) [8] que han tenido éxito en resolver problemas de ruteo de vehículos, corte de empaquetado (Strip Packing), entre muchos otros. Se han aplicado para el VRP original en [2] y [7].

Aplicando búsqueda en vecindarios variables (Variable Neighborhood Search) es referencia de su utilización el trabajo [6]. Para el caso de recocido simulado (Simulated Annealing), una metaheurística de trayectoria, es utilizada en [17], con resultados promisorios en un tiempo razonable para una variante de VRP con ventana de tiempo. Con búsqueda tabú (Taboo Search), otra metaheurística de trayectoria, en [13] se presenta una propuesta híbrida que combina la búsqueda tabú y recocido simulado obteniendo muy buenos resultados para las instancias analizadas en una variante de VRP. En una versión reciente en [3] para el VRP con una flota de vehículos heterogénea y distintas rutas aplica la búsqueda tabú y luego de realizados los experimentos con diversos benchmark del problema concluyen que su algoritmo produce soluciones de alta calidad en un tiempo computacionalmente aceptable. Utilizando colonias de hormigas (ACO) en [19] se presenta un algoritmo adaptativo con una búsqueda local Pareto aplicada a problemas de VRP y CVRP obteniendo resultados de mejor calidad comparados con metaheurísticas clásicas. Abordando enjambre de partículas (Particle Swarm) en [14]. Algunos ejemplos de algoritmos genéticos celulares se presentan en [1]. Una de las tendencias actuales es lograr obtener una formulación general para todos los problemas derivados del VRP, que los incluya como casos particulares. Un esfuerzo notable es el mostrado en [11] que se complementan el método heurístico y el algoritmo exacto unificado. Se observa al estudiar dichos trabajos, que el siguiente paso es lograr mejoras en el rendimiento de los algoritmos, ya sea mediante cambios en las estructuras de datos, o en cómo se acota el espacio de soluciones factibles o bien tratar de potenciarlos mediante la utilización de uno o más métodos, es decir, hibridándolos.

En este trabajo proponemos dos algoritmos evolutivos multirecombinativos que utilizan una mutación basada en conceptos de computación cuántica con el objetivo de mejorar la performance obtenida por el algoritmo sin hibridar.

El trabajo está organizado de la siguiente manera la Sección 2 describe el problema que se aborda en este trabajo. En la Sección 3 se introducen conceptos de hibridación y se muestran los algoritmos propuestos. La Sección 4 presenta el diseño de experimentos y los resultados obtenidos. Finalmente, la Sección 5 provee las conclusiones y futuras líneas de investigación.

2 Descripción del Problema

El VRP se puede definir como un problema de programación entera perteneciente a la categoría de problemas NP-duros. Entre las diferentes variedades de VRP trabajaremos con el VRP de Capacidad limitada (CVRP), en el que cada vehículo tiene una capacidad uniforme de un único artículo. Definimos el CVRP sobre un grafo no dirigido $G = (V, E)$ donde $V=\{v_0, v_1,..., v_n\}$ es un conjunto de vértices y $E = \{(v_i, v_j) / v_i, v_j \in V, i< j\}$ es un conjunto de ejes.

Los vértices v0 parten del depósito, y es desde donde m vehículos de capacidad Q deben abastecer a todas las ciudades o clientes, representados por un conjunto de n vértices $\{v_1, ..., v_n\}$.

Definimos E una matriz $C = (c_{ij})$ de costo, distancia o tiempo de viaje no negativos entre los clientes v_i y v_j. Cada cliente v_i tiene una demanda no negativa de artículos q_i

y tiempos de entrega δ_i (tiempo necesario para descargar todos los artículos). Siendo $v_1,..., v_m$ una partición de V, una ruta Ri es una permutación de los clientes en V_i especificando el orden en el que se visitan, comenzando y terminado en el deposito v_0. El costo de una ruta dada $R_i = \{ v_0, v_1, ..., v_{k+1} \}$, donde $v_j \in V$ y $v_0 = v_{k+1} = 0$ (*0* indica el depósito), viene dada por:

$$\mathrm{Cost}(\mathrm{R_i}) = \sum_{j=0}^{k} \mathrm{C_{i,\,j+1}} + \delta_{\mathrm{j}} \qquad \textbf{(1)}$$

y el costo de la solución al problema (S) es:

$$\mathrm{FCVRP}(\mathrm{S}) = \sum_{i=1}^{m} \mathrm{Cost}(\mathrm{R_i}) \qquad \textbf{(2)}$$

El CVRP consiste en determinar un conjunto de m rutas (i) de costo total mínimo - como especifica la ecuación (2); (ii) empezando y terminando en el deposito v0; de forma que (iii) cada cliente es visitado una sola vez por un sólo vehículo, sujeto a las restricciones (iv) de que la demanda total de cualquier ruta no exceda

$Q(\sum v_j \in R_i;\ \mathrm{q_j} \leq \mathrm{Q})$; y (v) la duración total de cualquier ruta no supera el límite preseleccionado $\mathrm{D}(\mathrm{Cost}(\mathrm{R_i}) \leq \mathrm{D})$. Todos los vehículos tienen la misma capacidad y transportan el mismo tipo de artículo. El número de vehículos puede ser un valor de entrada o una variable de decisión. En este estudio, la longitud de las rutas se minimiza independientemente del número de vehículos utilizados.

3 Hibridación de Metaheurísticas y Algoritmos Propuestos

En los últimos años, ha aumentado considerablemente el interés en las metaheurísticas híbridas en el campo de la optimización. Se han obtenido buenos resultados en muchos problemas de optimización clásicos y de la vida real utilizando metaheurísticas híbridas. Talbi en [15] y [16] propone una taxonomía para algoritmos híbridos y presenta dos clasificaciones para este tipo de algoritmos: jerarquizada y plana. Atendiendo a la taxonomía propuesta por Talbi, podemos decir que las hibridaciones propuestas en este trabajo se acercan a una hibridación de bajo nivel desde el punto de vista jerárquico y homogénea desde el punto de vista plano.

Para resolver el CVRP se utiliza el algoritmo MCMP-SRI [18] como algoritmo base, y dos hibridaciones que usan un operador de mutación basado en computación cuántica [5].

El algoritmo MCMP-SRI (Multiple Crossover Multiples Parents – Stud and Random Inmigrates) fue aplicado en diferentes problemas de planificación de máquina única para casos estáticos y casos dinámicos y los resultados obtenidos fueron satisfactorios. Además se lo utilizó el manejo de restricciones con metaheurísticas en [18] con resultados promisorios.

Para codificar las visitas a los clientes, que representan una posible solución, se utilizó una permutación de números enteros donde cada permutación $pi = (p1, p2, \ldots, p_n)$ es un cromosoma en el que pi representa el cliente i que debe ser visitado y n representa la cantidad de clientes a visitar. El cromosoma define el orden de la secuencia a seguir para visitar cada cliente, la función objetivo es minimizar la distancia total del plan de ruta general para satisfacer las demandas de todos los clientes, teniendo en cuenta la capacidad Q del vehículo. Se crea una población inicial de soluciones generadas aleatoriamente *(Stud(0))*, y luego estas soluciones son evaluadas. A continuación, la población pasa por un proceso multirecombinativo, donde el algoritmo crea un pool de apareamiento que contiene, padres generados aleatoriamente y un individuo semental, seleccionado de la población a través de selección proporcional. El proceso de crear descendientes funciona de la siguientre manera: el individuuo semental se aparea con cada uno de los padres. Las parejas se somenten a operaciones de recombinación y se generan 2 * n_2 (n_2 <= max_padres) descendientes. El mejor de los 2 * n_2 descendientes se almacena en un pool temporal de hijos. La operación de recombinación se repite n_1 veces (max_recombinaciones) hasta que el pool de hijos se complete. Finalmente, el mejor descendiente creado de n_2 padres y n_1 recombinaciones se inserta en la nueva población. El método de recombinación utilizado fue PMX (Partial Mapped Crossover) [9]: que puede verse como una extensión del cruzamiento de dos puntos para representaciones basadas en permutaciones. La selección de individuos fue a través de selección proporcional. En la Figura 1 se muestra el código del algoritmo MCMP-SRI.

```
MCMP-SRI

t=0; {Generacion Incial}
inicializar (Stud(t));
evaluar (Stud(t));
   while (not max_evaluaciones) do
   pool_aparemiento= Inmigrantes_generados_aleatoriamente ∪ Seleccionar(Stud(t));
                     while (not max_recombinaciones) do
                     evolucionar (pool_apareamiento);  {recombinación y mutación}
                     end while
   end while
   evaluar(pool_apareamiento);
   Stud(t+1) = seleccionar la nueva poblacion del pool_apareamiento
     t=t+1
   end while
```

Figura 1 Algoritmo MCMP-SRI

El algoritmo MCMP-SRI está basado en un AG y por lo tanto luego de la operación de crossover, el individuo pasa por una operación de mutación con una probabilidad muy baja. Los algoritmos propuestos usan una mutación basada en conceptos de computación cuántica, en particular qbit [5] y [10] donde se perturban los alelos de la siguiente manera: en el momento de realizar la mutación se elige una ventana de tamaño tres (este valor se seleccionó debido a que es un valor pequeño y controlado para la manipulación en el experimento). Se realizan todas combinaciones posibles para ese tamaño de ventana, evaluando en cada caso el individuo resultante y

quedándose con el mejor valor (mejor fitness). Al algoritmo resultante se lo denomina MCMP-SRI-MC y su funcionamiento se muestra en la Figura 2.

El siguiente algoritmo propuesto se denomina MCMP-SRI-MCV y en este caso para la mutación cuántica se utiliza una ventana de tamaño tres (igual que en el algoritmo anterior) pero los elementos que forman la ventana se eligen en forma aleatoria y en posiciones que no corresponden a una cadena adyacente en el cromosoma. Esto permite una mayor diversidad en la explotación. Luego de elegido los elementos que conforman la ventana se realizan todas las combinaciones posibles y se queda con la combinación que obtenga el mejor fitness. La Figura 3 muestra un ejemplo de este mecanismo.

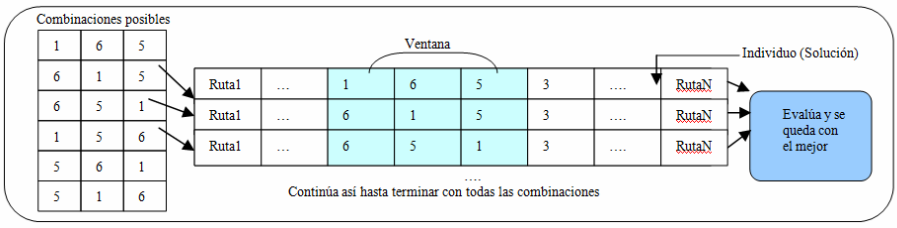


Figura 2 Mecanismo de MCMP-SRI-MC

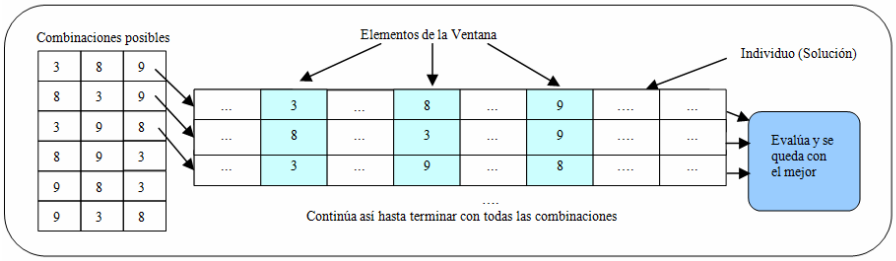


Figura 3 Mecanismo de MCMP-SRI-MCV

4 Diseño de experimentos y resultados

Para analizar la performance de los algoritmos propuestos se utilizaron las instancias provistas por Augerat et al.[1]. Se realizaron 30 corridas independientes de todos los algoritmos y se utilizaron 10 instancias (las más representativas) de Augerat. Se compararon los resultados de las versiones híbridas con los resultados obtenidos por un Algoritmo Genético Simple (AG) y MCMP-SRI.

La batería de problemas para éste trabajo está constituida por instancias, donde tanto las posiciones de los clientes como las demandas se generan aleatoriamente mediante una distribución uniforme. El tamaño de las instancias está en el rango de 31 a 79

[1] http://www.coin-or.org/SYMPHONY/branchandcut/VRP/data/index.htm.old

clientes. La parametrización común utilizada en los algoritmos es la siguiente: la población se generó de forma aleatoria y se fijó en 15 individuos. El criterio de parada se estableció en 5000000 evaluaciones. En método de crossover utilizado es PMX y la mutación es de intercambio (en AG y MCMP-SRI). Se estableció la probabilidad de mutación en 0,05 y la probabilidad de recombinación en 0,65. El número n_1 (número de operaciones de recombinación) y n_2 (número de padres) se estableció en 16 y 18 respectivamente, estos parámetros se seleccionaron en base a la experimentación de los valores previamente usados exitosamente [18].

Todos los algoritmos se implementaron en Java y las ejecuciones se realizaron en un procesador 2,53 GHz Intel i5 bajo Windows 7.

En la Tabla 1 se muestra el resumen de las 30 corridas independientes para las 10 instancias seleccionadas. Las dos primeras columnas corresponden a la instancia utilizada y a su valor óptimo. Luego, para cada uno de los algoritmos se muestra la mediana y el error porcentual con respecto al valor óptimo. Los mejores valores para la mediana y para el error porcentual se colocan en negrita. Podemos observar una clara diferencia en los resultados obtenidos entre el algoritmo AG y las versiones multirecombinativas. Además los algoritmos propuestos obtienen para todas las instancias mejores resultados con respecto a la mediana y el error porcentual comparado con MCMP-SRI. La propuesta MCMP-SRI-MCV obtiene en 6 de las 10 instancias los mejores valores para las variables de performance analizadas (mediana y error porcentual).

Tabla 1 Resultados obtenidos por AG, MCMP-SRI, MCMP-SRI-MC y MCMP-SRI-MCV para las instancias analizadas.

		AG		**MCMP-SRI**		**MCMP-SRI-MC**		**MCMP-SRI-MCV**	
Instancia	Optimo	Mediana	Error	Mediana	Error	Mediana	Error	Mediana	Error
A-n33-k5	661	863,72	0,31	706,20	0,07	695,88	0,06	**695,37**	**0,05**
A-n33-k6	742	895,87	0,21	770,61	**0,04**	**765,43**	**0,04**	771,39	**0,04**
A-n34-k5	778	970,02	0,24	833,41	0,07	823,03	0,06	**813,21**	**0,05**
A-n37-k6	949	1139,65	0,21	1004,87	0,07	999,46	0,06	1005,42	**0,06**
A-n45-k7	1146	1396,02	0,22	1235,81	0,08	1233,50	0,07	1233,67	0,08
A-n53-k7	1010	1467,17	0,46	1177,63	0,16	1174,63	0,17	**1166,48**	**0,15**
A-n61-k9	1035	1449,55	0,41	1214,39	0,16	**1201,36**	**0,15**	1204,55	0,16
A-n62-k8	1290	1840,78	0,42	1487,72	0,16	1485,29	0,16	**1474,55**	**0,14**
A-n64-k9	1402	1881,99	0,35	1600,69	0,14	1596,97	0,14	**1572,27**	**0,12**
A-n80-k10	1764	2553,64	0,45	2072,06	0,18	2092,03	0,20	**2049,61**	**0,16**

Para realizar un análisis más profundo de los resultados únicamente se analiza el error porcentual. Comenzaremos verificando las condiciones necesarias para aplicar test estadísticos paramétricos o no paramétricos, estas condiciones son: Independencia, Normalidad y Homocedasticidad. Como los resultados a analizar provienen de corridas independientes sólo debemos verificar las otras dos condiciones. Para todos los test utilizados se obtiene el p-valor asociado, que representa la disimilitud de los resultados de la muestra con respecto a la forma normal. Por lo tanto, un p-valor bajo, señala una distribución no-normal. En este estudio consideramos un nivel de significancia $\alpha = 0,05$ por lo tanto un p-valor mayor que α indica que se cumple la condición de normalidad. Se utilizó como herramienta estadística SPSS.

En la Tabla 2 se muestran los resultados obtenidos por los test aplicados. El resultado del primer test aplicado, test de Kolmogarov-Smirnov, está compuesto por las cuatro columnas que representan los algoritmos analizados.
El símbolo "*" indica que los resultados no cumplen con la condición de normalidad y el valor a continuación representa el p-valor en cada caso.
En la ante-última columna se muestra el resultado de la aplicación del test de Levene para analizar la homocedasticidad. Para este caso el símbolo "*" indica que los resultados no cumplen la condición de homocedasticidad. La última columna representa el resultado de la aplicación del test de Kruskal-Wallis.

Tabla 2 Test de normalidad de Kolmogorov-Smirnov, Test de Levene y Test de Kruskal-Wallis.

	Test Kolmogorov-Smirnov				Test de Levene	Test de Kruskal-Wallis
Instancias	AG	MCMP-SRI	MCMP-SRI-MC	MCMP-SRI-MCV		
A-n33-k5	0,20	0,20	0,20	0,54	0,00*	(+)
A-n33-k6	0,20	0,03*	0,06	0,16*	0,00*	(+)
A-n34-k5	0,20	0,20	0,20	0,12	0,00*	(+)
A-n37-k6	0,20	0,06	0,19	0,19*	0,00*	(+)
A-n45-k7	0,13	0,20	0,19	0,20	0,00*	(+)
A-n53-k7	0,20	0,20	0,20	0,20	0,00*	(+)
A-n61-k9	0,20	0,20	0,20	0,20	0,00*	(+)
A-n62-k8	0,20	0,13	0,20	0,20	0,00*	(+)
A-n64-k9	0,20	0,20	0,20	0,20	0,00*	(+)
A-n80-k10	0,20	0,20	0,12*	0,20	0,00*	(+)

Teniendo en cuenta que no se cumplen las condiciones para realizar los test paramétricos se aplica entonces el test de Kruskal-Wallis para determinar si existen diferencias significativas entre los algoritmos. Utilizamos el signo (+) para especificar que existen diferencias significativas entre los resultados obtenidos por los algoritmos y (-) en caso contrario. Podemos observar que en todos los casos las diferencias entre los resultados obtenidos por los algoritmos son estadísticamente significativas.

Ahora para saber entre los resultados de qué algoritmos existen diferencias estadísticamente significativas aplicamos el test de Tukey. Luego se la aplicación de este último test únicamente se encontraron diferencias estadísticamente significativas entre el algoritmo genético (AG) y los versiones multirecombinativas.

En las Figuras 5 y Figura 6 se muestra cómo se distribuyen los resultados con respecto a la mediana y podemos notar claramente la diferencia de resultados obtenidos por el algoritmo AG y los restantes algoritmos. En la Figura 5, para la instancia A-n64-k9, vemos que MCMP-SRI-MCV obtiene en mediana el menor error porcentual pero MCMP-SRI-MC es más robusto. En la Figura 6, para la instancia A-n53-k7, podemos observar que los valores obtenidos por MCMP-SRI-MCV son levemente menores a los valores obtenidos por los otros dos algoritmos multirecombinativos. Si bien el algoritmo es más robusto, no obstante, estas diferencias no son estadísticamente significativas.

Con la idea de analizar ahora el desempeño de los algoritmos sobre el conjunto de problemas aplicaremos un test no paramétrico (para muestras relacionadas). El siguiente test se ha aplicado utilizando el paquete (en Java) CONTROLTEST que se

puede obtener en el sitio web público temático SCI2S - Statistical Inference in Computational Intelligence and Data Mining.

Aplicamos el Test de Friedman Alineado, que se trata de un equivalente no paramétrico del test de ANOVA. Se basa en n conjuntos de filas, donde el rendimiento de los algoritmos analizados se clasifica por separado para cada conjunto de datos. En esta técnica, un valor de posición se calcula como el promedio de rendimiento conseguido por todos los algoritmos en cada conjunto de datos. Luego, se calcula la diferencia entre el rendimiento obtenido por un algoritmo y el valor de la ubicación. Este paso se repite para algoritmos y conjuntos de datos. Las diferencias resultantes, llamadas observaciones alineadas, que mantienen su identidad con respecto al conjunto de datos y la combinación de algoritmos a los que pertenecen. Seguidamente, se clasifican de 1 a k_n los algoritmos.

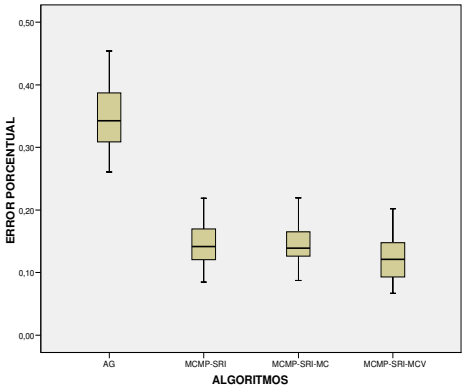


Figura 5 Ejemplo de Diagrama de Cajas (Boxplot) para la instancia A-n64-k9

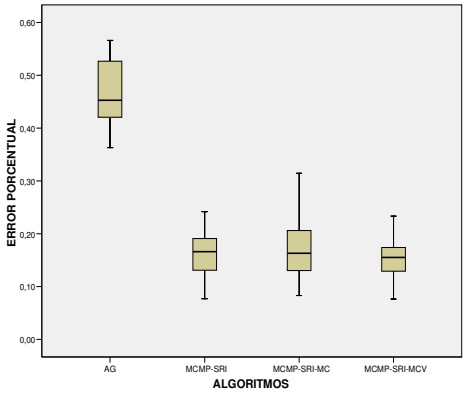


Figura 6 Ejemplo de Diagrama de Cajas (Boxplot) para la instancia A-n53-k7

Los resultados del estadístico de Friedman Alineado (distribuido de acuerdo a χ^2 con 4 grados de libertad: 7,76) son: AG ranking 3,89; MCMP-SRI ranking 2,28; MCMP-SRI-MC ranking 2,12; MCMP-SRI-MCV ranking 1,72. El valor de p calculado por el test de Friedman Alineado es: 0,10 (valor > 0,05) esto significa que existen diferencias estadísticamente significativas entre los resultados obtenidos por algoritmos. Podemos observar que nuestra propuesta algoritmo MCMP-SRI-MCV es el que obtiene el mejor ranking luego le sigue MCMP-SRI-MC, continua MCMP-SRI y el peor ranking es el obtenido por el AG.

5 Conclusiones

Los algoritmos evolutivos son algoritmos de búsqueda robustos en el sentido que proporcionan buenas soluciones en una amplia clase de los problemas que de otra manera serían computacionalmente intratables. Para mejorar el funcionamiento de los AEs, se pueden usar enfoques multirecombinativos los cuales permiten múltiples

intercambios de material genético entre múltiples padres y con ello mejorar la velocidad de convergencia. Para enriquecer el proceso de búsqueda, mediante un mejor equilibrio entre la exploración y la explotación, el concepto de stud e inmigrantes aleatorios fue insertado en MCMP-SRI. La presencia del *stud* asegura la retención de los rasgos buenos de soluciones anteriores y los inmigrantes aleatorios, como una fuente continua de diversidad genética, evita la convergencia prematura.

El Problema de Ruteo de Vehículos con Capacidad limitada (CVRP), es una de las variantes del VRP que es considerado emblemático en el campo de la distribución, logística y tráfico.

En este trabajo hemos presentado dos versiones de MCMP-SRI que utilizan una mutación basada en conceptos de computación cuántica (MCMP-SRI-MC) y (MCMP-SRI-MCV). Hemos aplicado una familia de test no paramétricos para comparar los resultados (error porcentual) de los algoritmos y se realizaron comparaciones múltiples.

En cuanto a los resultados obtenidos, luego de aplicar el test no-apareado (Kruskal-Wallis) podemos afirmar que existen diferencias estadísticamente significativas entre los algoritmos en cada una las instancias analizadas. Al aplicar el test post-hoc (Tukey) confirmamos que las diferencias correspondían a los enfoques multirecombinativos con respecto al algoritmo AG. Seguidamente aplicamos un test apareado (Friedman Alineado). En Friedman Alineado, obtuvimos diferencias estadísticamente significativas en los resultados de los algoritmos. Con este test el algoritmo MCMP-SRI-MCV obtiene los mejores resultados y las diferencias son estadísticamente significativas. Trabajos futuros incluirán otras formas de hibridación y su aplicación a otras variantes de VRP.

Agradecimientos

Se agradece la cooperación del equipo de proyecto del LabTEm y a la Universidad Nacional de la Patagonia Austral, de los cuales se recibe apoyo continuo. El último autor agradece el constante apoyo de la Universidad Nacional de San Luis y a ANPCYT que financia su investigación.

Referencias

1. Alba, E. y Dorronsoro B.; Solving the Vehicle Routing Problem by Using Cellular Genetic Algorithms. Evolutionary Computation in Combinatorial: pp.1-10. (2004).
2. Bäker, B.M. y Ayechew M.A. A genetic algorithm for the vehicle routing problem. Computers & Operations Research, pp. 787-800. (2003).
3. Brandão, J. A tabu search algorithm for the heterogeneous fixed fleet vehicle routing problem. Computers & Operations Research, 38(1), 140-151.(2011).
4. Christofides, N., Mingozzi A. y Toth P. The vehicle routing problem. Reveu francaise d´automatique d´informatique et de recherche opérationnelle. vol.1.tome10 (Combinatorial Optimization).pp.315–338.(1979).
5. Deutsch, D. Quantum theory, the church-turing principle and the universal quantum computer. Proceedings of the Royal Society of London A.400:97–117. (1985).

6. Fleszar, K. y Osman, I. H. A variable neighborhood search algorithm for the open vehicle routing problem. European Journal of Operational Research vol.195: pp.803-809. (2009).
7. Goel, A. y Gruhn, V. A General Vehicle Routing Problem. European Journal of Operational Research vol.191: 650–660. (2008).
8. Goldberg, D. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Co. (1989).
9. Goldberg, D y Lingle, R. Alleles, loci and the traveling salesman problem. Proc. of the First International Conference on Genetic Algorithms, Lawrence Eribaum Associates, pp.154-159.Hilldale, NJ. (1987).
10. Han, K y Kim, J. Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. IEEE Trans. Evolutionary Computation, 6:580–593. (2002).
11. Huang S., Fu X., Chen P., Ge CC y Teng S. An Application Study on Vehicle Routing Problem Based on Improved Genetic Algorithm. Journal of Pervasive Computing and the Networked World. Lecture Notes in Computer Science vol. 7719. pp. 246-258. (2013).
12. Laporte, G. The vehicle routing problem: An overview of exact and approximate algorithms. European Journal of Operational Research vol.59 pp.345-358. (1992).
13. Lin S.W., Ying K. C., Lee C. y Lee Z.J.A Hybrid Approach for Vehicle Routing Problem with Time Windows. Journal of Computational Information Systems 7: 13 pp. 4939-4946. (2011).
14. Liu J., y Kachitvichyanukul, V. A New Solution Representation for Solving Location Routing Problem via Particle Swarm Optimization. Proceedings of the Institute of Industrial Engineers Asian Conference 2013. pp 103-110. (2013).
15. Talbi, E.-G. A taxonomy of hybrid metaheuristics. Heuristics, pp.541–564. (2002).
16. Talbi, E.-G.. Metaheuristics: From design to Implementation. Wiley. (2009).
17. Tavakkoli-Moghaddama R., Gazanfarib M., Alinaghianb M., Salamatbakhshc A. y Norouzic N. A new mathematical model for a competitive vehicle routing problem with time windows solved by simulated annealing. Journal of Manufacturing Systems. vol 30(2),pp. 83–92. Technical paper. (2013).
18. Villagra A. , Pandolfi D. y Leguizamón G. Handling constraints with an evolutionary tool for scheduling oil wells maintenance visits. Engineering Optimization. vol. 45(8), pp. 963-981. (2013).
19. Wang Y.P. Adaptive Ant Colony Algorithm for the VRP Solution of Logistics Distribution. Research Journal of Applied Sciences, Engineering and Technology 6(5):807-811, 2013ISSN:2040-7459. Maxwell Scientific. (2013).

Análisis del comportamiento de un AG para GPUs

Carlos Bermúdez, Carolina Salto

Facultad de Ingeniería - Universidad Nacional De La Pampa
bermudezc@yahoo.com, saltoc@ing.unlpam.edu.ar
http://www.ing.unlpam.edu.ar

Resumen Este trabajo presenta un algoritmo genético simple ejecutándose sobre GPU y empleando la tecnología CUDA para resolver el problema MaxCut. Se realiza un estudio empírico del impacto en el rendimiento del algoritmo en la elección de distintos operadores de cruce para representaciones binarias. Las pruebas mostraron un buen desempeño de las distintas variantes planteadas, aunque una mejor calidad de resultados se obtuvo con la variante utilizando un cruce de dos puntos de corte. El paso siguiente fue contrastar el rendimiento de este algoritmo con una misma versión pero ahora ejecutándose en serie sobre CPU y así poder determinar la ganancia de tiempo, reflejada por el *speedup*. Los resultados obtenidos indican que la ganancia en tiempo está relacionada con la densidad del grafo que representa cada instancia del MaxCut.

Keywords: Algoritmo genético, GPU, CUDA, MaxCut

1. Introducción

Los algoritmos genéticos (AG) desarrollados por Holland en 1975 [1] son una herramienta muy potente cuando de optimizar se trata. Y sus beneficios están ampliamente demostrados en la bibliografía existente. Los AG son algoritmos de búsqueda estocásticos basados en población y para obtener buenos resultados se necesita, a menudo, crear y evaluar muchas soluciones candidatas. Por otra parte, estos algoritmos son inherentemente paralelos debido a que una solución candidata puede crearse o evaluarse de forma independiente a las demás. Por otra parte, las Unidades de Procesamiento Gráfico (GPU) ofrecen un gran poder de cómputo a precios accesibles. Estas unidades, que pueden ser fácilmente programadas, a través de una extensión del lenguaje C (CUDA (Compute Unified Device Architecture) [4]), permiten implementar programas que hagan uso de las capacidades de las GPU. A pesar de que estos chips fueron diseñados para acelerar la rasterización de primitivas gráficas, su rendimiento ha atraído a una gran cantidad de investigadores que los utilizan como unidades de aceleración para muchas aplicaciones científicas [2]. Esto da lugar a que el uso de las GPU como soporte para la ejecución de AGs se esté popularizando. En comparación con una CPU, las GPU más recientes son unas 15 veces más rápidas que los procesadores de seis núcleos de Intel en operaciones de punto flotante de precisión

simple [12]. Dicho de otra manera, un grupo con una sola GPU por nodo ofrece un rendimiento equivalente a un cluster de 15 nodos de un sólo CPU.

En este trabajo, diseñamos un AG que se ejecuta completamente sobre una GPU (AG-GPU) para resolver el problema de corte máximo de un grafo, conocido como MaxCut. Es un problema NP-duro muy conocido y además de su importancia teórica, tiene aplicaciones en varios campos y ha sido reformulado de varias maneras. El objetivo de este trabajo es medir el desempeño del AG diseñado para ejecutar en CPU teniendo en cuenta dos tipos de crossover (un punto y dos puntos), para tratar de determinar si el beneficio obtenido en la calidad de los resultados con un operador de dos puntos es más conveniente que utilizar la versión con un solo punto de corte. Por último, se calcula el speed-up del algoritmo, para lo cual se ejecutó el algoritmo en su versión serie sobre CPU y con esto poder contrastar los tiempos obtenidos.

El trabajo se organiza de la siguiente manera. La Sección 2 introduce el problema de optimización utilizado MaxCut y la descripción del algoritmo genético utilizado en la experimentación. En la Sección 3 se detalla la parametrización utilizada en la experimentación y en la Sección 4 el análisis de los resultados obtenidos. Finalmente la Sección 5 presenta las conclusiones alcanzadas y la propuesta de trabaja futuro.

2. Algoritmo Genético para GPU

Esta sección describe los detalles de implementación del AG para GPU utilizado en el trabajo para resolver el problema de corte máximo de un grafo, denominado MaxCut. Por ello esta sección se estructura en dos partes: la primera está dedicada a la descripción del problema y la segunda a introducir detalles de implementación del algoritmo AG-GPU desarrollado para resolver tal problema.

2.1. Problema MaxCut

El problema MaxCut [5] es un problema que pertenece a la clase de los NP-Completos, y se define de la siguiente manera:

Definition 1. *Dado un grafo no dirigido G=(V,E) y pesos no negativos $w_{ij} = w_{ji}$ en los arcos $(i, j) \in E$, el problema consiste en encontrar un conjunto de vértices S que maximicen el peso de los arcos en el corte $(S,\overline{S})$, esto es, el peso de los arcos que tienen un vértice en S y el otro en $\overline{S}$.*

El peso del corte queda denotado por $w(S, \overline{S}) = \sum i \in S, j \notin Sw_{ij}$, y como se menciona en la definición, dicho peso tendrá que ser máximo.

2.2. Descripción del algoritmo AG-GPU

Aprovechando las ventajas brindadas por la GPU para acelerar el proceso de optimización, se implementaron todos los operadores del AG-GPU para que se ejecuten como *Kernel* en la GPU. Por lo tanto, el bucle principal de este AG

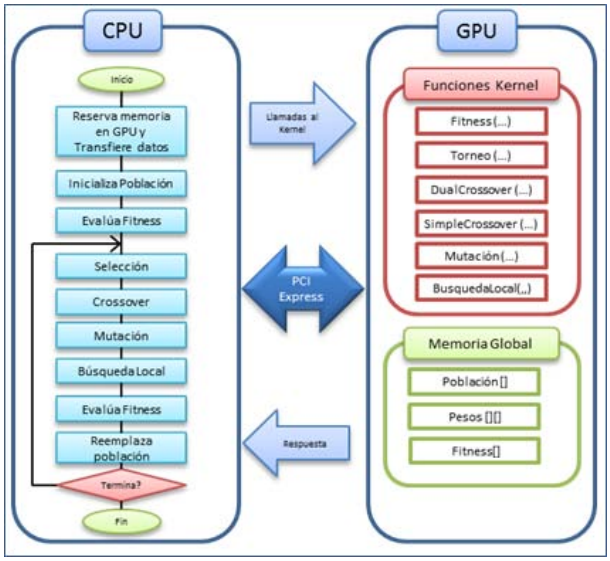


Figura 1. Configuración del AG en CPU y GPU con CUDA

se implementó para que corra completamente en el *host* y todos los operadores genéticos se implementaron para ser ejecutados en la GPU. En la Figura 1 se puede observar esta configuración. El primer paso para trabajar con la GPU es reservar la memoria necesaria donde se alojarán los arreglos utilizados. Es importante resaltar que la GPU no posee un generador de números aleatorios, por lo tanto se debe crear un vector en la CPU, llenarlo con los valores deseados y luego transferirlos a la GPU. De la misma forma se debe enviar como parámetro la posición que se utilizará de este arreglo. Para la representación de la población se utilizó codificación binaria basada en nodos, es decir que si un bit está en uno significa que el nodo que está representando ese bit forma parte del conjunto S. La representación binaria es preferible a trabajar directamente con enteros ya que, a pesar de ser un poco más complejo su manejo, utiliza mucho menos memoria. El uso de la memoria que se utiliza en el *device* es crítico, ya que se tiene que estar intercambiando información entre el *host* y el *device* continuamente.

Para almacenar un individuo se utilizó el tipo de dato unsigned char que representa la mínima pieza de información denominada palabra. Por ejemplo, un individuo que represente una instancia de 10.000 nodos, serían necesarias 1.250 palabras del tipo unsigned char para ser almacenado. Se debe contemplar también, que si el tamaño de la instancia no es múltiplo del tipo de dato van a existir datos inválidos al final de la última palabra. Para inicializar cada individuo se genera cada palabra con ceros y unos de forma aleatoria. Al evaluar el fitness del individuo se deben recuperar sus bits uno por uno, para esto, se debe utilizar una máscara auxiliar (2unsigned char -1) y se efectúa un AND con la palabra que se está evaluando, luego se desplaza a la derecha la máscara con el

operador >> y se vuelve a comparar, así hasta el final del individuo. Cuando un bit del individuo tiene que ser mutado, se necesita primero saber en qué palabra se encuentra y la posición que ocupa en la palabra. Luego se debe desplazar hacia la derecha la máscara hasta la misma posición de este bit y se efectúa un XOR entre la máscara y la palabra seleccionada.

Tanto para el operador de torneo como el de mutación se lanza un thread por cada individuo. En el operador de torneo binario cada thread evalúa el fitness de los individuos involucrados en el torneo y entonces almacena la posición del ganador en la memoria global para que luego pueda ser utilizado por el operador de crossover. En la mutación, cada thread decide cuándo el individuo tiene que ser mutado, utilizando para esto, un valor random de la memoria global. Si el individuo tiene que ser mutado se utiliza otro valor random de la memoria para ubicar el bit a alterar.

Como el objetivo del trabajo se centra en analizar cual operador de crossover es el más adecuado para resolver el problema, es que se consideraron dos operadores: el operador de un punto y el operador de dos puntos. Además como es sabido dentro de la comunidad de computación evolutiva , el rendimiento del algoritmo está fuertemente vinculado con el tipo de operador seleccionado [13].

El crossover de un punto (SPX) selecciona un punto en el vector de los padres. Todos los bits más allá de este punto se intercambiarán entre los dos padres. Para implementar el SPX se necesita calcular en qué palabra y en qué posición de ésta se hará el corte. Si la posición es al comienzo de la palabra simplemente se copiarán las palabras completas y no son necesarias operaciones a nivel de bit. En caso contrario se necesitan operaciones a nivel de bit en la palabra seleccionada y las palabras restantes se copian de forma directa. La idea del algoritmo es desplazar los bits hacia un lado para descartar la información innecesaria y luego desplazarlos en sentido contrario para dejar los bits requeridos en la ubicación correcta.

El crossover de dos puntos (DPX) requiere seleccionar dos puntos en los vectores de los padres. Todos los datos entre los dos puntos se intercambian entre los padres, creando dos hijos. El crossover de dos puntos es similar al SPX pero es necesario contemplar situaciones especiales. En el caso más general, los dos puntos de corte caen dentro de palabras distintas y se debe aplicar el mismo procedimiento que SPX para ambas palabras. Por otra parte cuando los dos puntos de corte caen dentro de la misma palabra las operaciones de desplazamiento para obtener el bit requerido se deben efectuar sobre la misma palabra. Otro caso especial que requiere atención, sobre todo para ganar eficiencia, es cuando los dos puntos de crossover caen en la misma palabra y al comienzo de ésta.

El operador de crossover sobre GPU se organizó en forma de bloques de threads, la cantidad de bloques es igual a la mitad de la población. Cada thread maneja varios elementos de la solución, para esto toma el identificador de bloque *blockid* al que pertenece y copia el primer bit en la posición *threadid*, luego el segundo bit en la misma posición del segundo bloque y así hasta completar la solución. Para mejorar el rendimiento del operador se utilizó acceso coalescente

a memoria global, así los threads contiguos acceden a localizaciones de memoria adyacentes.

Después de aplicar los operadores genéticos de crossover y mutación, los nuevos individuos creados pasan a un proceso de optimización local con una cierta probabilidad. Este algoritmo es una variación eficiente de la fase de búsqueda local propuesta por Festa et al. [6] . Cuando es aplicado sobre una solución x, el procedimiento comienza creando un conjunto T de todos los nodos i. El procedimiento iterativo selecciona de forma aleatoria un nodo i del conjunto T y entonces cambia este nodo desde un subconjunto al otro en x de acuerdo a las siguientes reglas:

$$\textbf{If} \quad (xi = 0 \, and \, \sigma s(i) - \sigma \overline{s}(i) > 0) \quad \textbf{then} \quad xi = 1 \tag{1}$$

$$\textbf{If} \quad (xi = 1 \, and \, \sigma \overline{s}(i) - \sigma s(i) > 0) \quad \textbf{then} \quad xi = 0 \tag{2}$$

Donde, para cada nodo $i : 1, \ldots, n$; $\sigma s(i) = \sum_{j \in S} w_{ij}$ y $\sigma \overline{s}(i) = \sum_{j \in \overline{S}} w_{ij}$ representan el cambio en la función objetivo asociada con mover un nodo i desde un subconjunto del corte al otro. Todos los posibles movimientos de una solución son investigados simplemente haciendo un intercambio entre todos los nodos. La solución actual es reemplazada por la mejorada. El seudocódigo del procedimiento de búsqueda local se puede observar en el algoritmo 1.

Algoritmo 1: Heurística de búsqueda local

```
T ← V ;
while T ≠ 0 do
    i ← verticealeatorioenT ;
    if (i ∈ s)and(σ_s(i) − σ_s̄(i) > 0) then
        S ← S\ {i} ;
        S̄ ← S̄∪ {i} ;
    end
    if (i ∈ s̄)and(σ_s̄(i) − σ_s(i) > 0) then
        S̄ ← S̄\ {i} ;
        S ← S∪ {i} ;
    end
    T ← V{i} ;
end
```

Para el funcionamiento de este operador sobre GPU se lanza un thread por cada individuo en donde se evalúa el beneficio de cambiar un nodo de subconjunto, de ser positivo el intercambio se incorpora a la solución y se continua evaluando el siguiente nodo hasta terminar con el cromosoma.

3. Experimento

Para realizar la experimentación se utilizó el conjunto de instancias generadas por Helmberg y Rendl [7]. Éstas consisten en grafos toroidales planos y generados aleatoriamente variando el tamaño y la densidad, con pesos que toman el valor 0, 1 o -1. El tamaño del grafo V varía de 800 a 2000 nodos. La densidad fluctúa

desde 0.25 % a 6.12 % como se observa en la Tabla 1. Festa et al. [6], Martí et al.[8] , y Arráiz et al [9] usan estos grafos en sus experimentos, por lo tanto es una elección conveniente para propósitos comparativos.

Cuadro 1. Cantidad de nodos y densidad para cada instancia.

Instancias	—V—	Densidad (%)
G1, G2, G3	800	6.12
G11, G12, G13	800	0.63
G14, G15, G16	800	1.58
G22, G23, G24	2000	1.05
G32, G33, G34	2000	0.25
G43, G44, G45	1000	2.10

Para seleccionar el valor correspondiente a cada uno de los parámetros que controlan a los operadores genéticos, se efectuaron diferentes pruebas con las instancias G1 y G22 (se las consideraron representativas del conjunto ya que poseen distintas cantidades de nodos y de densidad). Estas pruebas consistieron en variar los siguientes parámetros: probabilidad de mutación, probabilidad de crossover, tamaño de la población, cantidad de generaciones y probabilidad de la búsqueda local. Para cada una de estas pruebas se efectuaron 30 ejecuciones y se tomaron los promedios. Los parámetros utilizados fueron: población de 200 y 480 individuos, 100, 200 y 300 generaciones y las probabilidades de los operadores genéticos tomaron valores de 60 %, 80 % y 100 %. Los resultados de estas pruebas sugirieron que los siguientes valores son los que arrojaron la mejor calidad de resultados: tamaño de población igual a 480 individuos, 200 generaciones, probabilidad de mutación igual al 100 % (todos los individuos se mutan y cada gen tiene una probabilidad 1/—V— de ser cambiado), probabilidad de crossover del 60 % y probabilidad de la búsqueda local igual al 100 % (por motivos de espacio no se incluyen las tablas con estos resultados). Estos resultados fueron sometidos a una validación estadística utilizando tests no paramétricos con un valor de significancia del 95 %. Si bien una probabilidad del 100 % para la búsqueda local puede parecer extraña, este mecanismo ayuda al algoritmo a converger en forma más temprana sin caer en óptimos locales.

Los resultados mostrados en la siguiente Sección son el promedio de 30 ejecuciones independientes. Se efectuó un análisis de estadístico por medio del test no-paramétrico Kruskal Wallis para distinguir las diferencias más relevantes a través de la media de los resultados para cada algoritmo. Se consideró un nivel significativo de $\alpha = 0.05$, para indicar un nivel de confianza del 95 % en los resultados.

Los algoritmos fueron implementados en C++ utilizando la librería CUDA, y fueron ejecutados en una máquina con un microprocesador Intel© Core™ i7 2600 CPU @ 3.40GHZ con 4GB de RAM y una placa de video G-FORCE GTX-580 con 512 cores y 3 GB de memoria.

Cuadro 2. Resultados de las variantes de AG-GPU.

	AG-GPU_SPX			**AG-GPU_DPX**			**test**
	Mejor	**Media**	σ	**Mejor**	**Media**	σ	
G1	11624	11561.77	± 25.82	11624	11589.10	± 19.63	+
G2	11605	11557.87	± 15.22	**11616**	11588.83	±11.95	+
G3	11602	11568.97	± 15.15	**11617**	11595.60	± 12.82	+
G11	**556**	547.67	± 3.83	520	498.33	± 11.03	+
G12	514	496.80	±10.30	**546**	537.33	± 4.18	+
G13	550	527.73	± 8.51	**572**	560.87	±5.35	+
G14	**3058**	3049.43	± 3.80	3048	3034.77	± 5.50	+
G15	**3036**	3028.43	±3.52	3032	3016.53	± 7.23	+
G16	**3039**	3032.80	±3.40	3033	3023.07	± 5.67	+
G22	13248	13187.80	±24.75	**13309**	13268.27	± 18.14	+
G23	13259	13192.23	±30.21	**13309**	13273.47	± 14.14	+
G24	13241	13184.93	±30.58	**13301**	13265.60	± 20.14	+
G32	1234	1187.67	±17.68	**1336**	1322.93	± 8.17	+
G33	1188	1155.73	±16.31	**1318**	1296.93	±11.55	+
G34	1200	1155.67	±17.97	**1314**	1295.80	±9.21	+
G43	6630	6601.30	± 14.56	**6649**	6625.43	± 12.41	+
G44	6626	6592.10	±16.01	**6644**	6626.20	±10.20	+
G45	6641	6596.70	±17.71	**6644**	6622.17	±13.75	+

4. Análisis de resultados

Esta sección está organizada en dos partes. La primera está orientada a determinar cuál variante del AG-GPU es la que mejor rendimiento obtiene. La segunda parte muestra una comparativa en el rendimiento del AG-GPU y su versión secuencial ejecutada en serie en CPU.

En el Cuadro 2 se puede observar el comportamiento de las dos variantes del AG-GPU para cada una de las instancias del MaxCut. En el Cuadro se muestra la siguiente información: mejor solución encontrada por el AG-GPU en las 30 ejecuciones (columna Mejor), promedio de las mejores soluciones (columna Media) junto con el desvío estándar (columna σ). En negrita se resaltan las mejores soluciones para cada instancia. Comparando las mejores soluciones del AG-GPU_DPX y AG-GPU_SPX, se puede distinguir el buen desempeño del AG-GPU_DPX, ya que en 14 de las 18 instancias obtuvo mejores resultados que AG-GPU_SPX. Este mejor desempeño se hace más evidente en las instancias que tienen una menor complejidad, como son las instancias G32, G33 y G34 que poseen una densidad de 0,25 % y el mejor valor encontrado supera en más de 100 unidades al de AG-GPU_SPX. Por el contrario, en las instancias con mayor densidad como las G1, G2 y G3, que tienen una densidad de 6,12 % , la ganancia que se obtiene está en el orden de las 10 unidades. Por último, no parece haber una relación directa entre la cantidad de nodos y la calidad de los resultados, ya quelos mejores resultados se obtuvieron en instancias que tienen 2000 nodos, mientras que en las que se arrojaron valores menores tienen 800 nodos. Por otra parte, podemos observar que existen diferencias estadísticamente significativas entre las dos variantes de AG-GPU, ya que los respectivos valores de p para la prueba estadística es mayor que el nivel de confianza de 0,05 (símbolo + en la columna test).

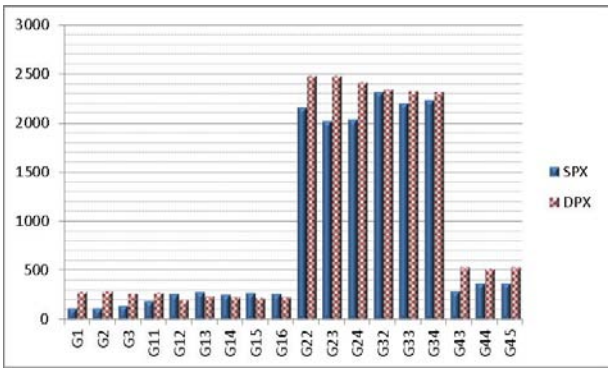


Figura 2. Comparación de los tiempos de ambos operadores.

En las Figuras 2 y 3 se puede observar los tiempos y las evaluaciones que consumieron las distintas variantes de AG. En general, la variante AG-GPU_DPX insume más tiempo y cantidad de evaluaciones que AG-GPU_SPX, excepto en las instancias G12, G13, G14, G15 y G16. Estas últimas también se corresponden con las que AG-GPU_DPX obtuvo peores resultados. Estas diferencias fueron corroboradas con los estudios estadísticos realizados.

En resumen, la variante AG-GPU_DPX fue la que mejor comportamiento ha exhibido (en 14 de las 16 instancias analizadas), por lo que se utilizará para realizar la siguiente experimentación.

Otro de los ejes de este trabajo es medir la ganancia de tiempo que se obtiene de ejecutar el algoritmo en paralelo (ambiente GPU) respecto de su ejecución en serie (entorno CPU). Para ello se efectuó la misma experimentción y con los mismos parámetros utilizados en la versión paralela, pero ahora con un algoritmo modificado para que pueda correr completamente en un CPU, esta implementación es con un único hilo de ejecución. Con estos resultados se pueden relacionar los mejores tiempos de la versión paralela con los mejores tiempos de su versión serie, y así obtener una medida de rendimiento conocida como *speed-up* [12]. En la Figura 4 se muestran los valores obtenidos para esta métrica para cada una de las instancias. En este caso se observa que el valor de *speed-up* está muy relacionado con la densidad que exhibe cada una de las instancias. Por ejemplo, el *speed-up* es mucho mayor a medida que aumenta la densidad de las instancias, las instancias G32, G33 y G34 que presentan una densidad de 0,25 % se obtiene un valor de *speed-up* de 1.18 mientras que en las instancias más densas G1,G2 y G3 con 6,12 % de densidad, el valor de *speed-up* es de 17, aproximadamente. Si bien los valores de *speed-up* no son importantes, vale destacar que la comparación se realiza con un procesador de última generación muy rápido.

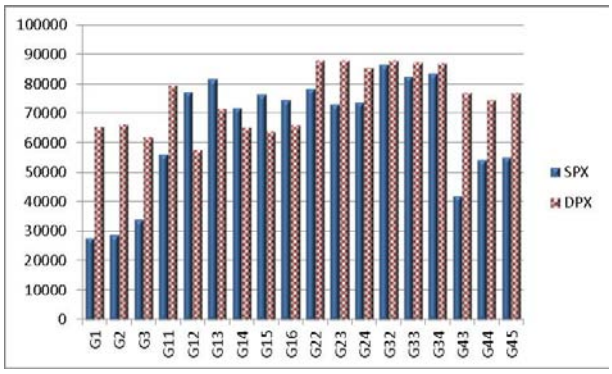


Figura 3. Comparación de las evaluaciones de ambos operadores.

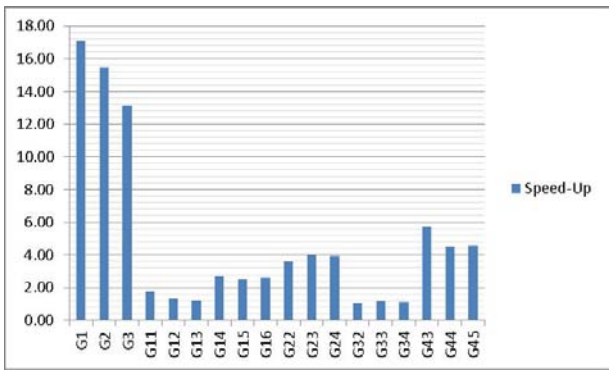


Figura 4. Speed-Up

5. Conclusiones

En este trabajo hemos presentado una implementación de un AG para GPU (AG-GPU) para resolver el problema de MaxCut. Para implementar el AG para GPUs se utilizó la tecnología CUDA, obteniéndose conocimiento sobre el manejo de AGs y su optimización en estos entornos. Se analizó el comportamiento del AG para GPUs con dos tipos de crossover: de uno (SPX) y dos (DPX) puntos. Hemos analizado el rendimiento del algoritmo utilizando 16 instancias del problema MaxCut. La variante AG-GPU_DPX mejora notablemente la calidad de las soluciones obtenidas por AG-GPU_SPX. El speedup de nuestra implementación fue investigada usando una tarjeta GeForce GTX 580 y un procesador rápido Intel Core i7 920 a 3.4 GHz. Para nuestra opción los valores de *speed-up* varían entre 1 y 17, y están relacionados con la densidad del grafo de corte que representa cada instancia. Una de las ventajas de nuestra implementación es que se puede ejecutar en cualquier GPU nVidia y plataforma Linux/Windows.

Nuestro trabajo futuro estará enfocado sobre estrategias de optimización específicas de CUDA para mejorar de ser posible los valores de speedup. También nuestra investigación estará orientada a la implementación de algoritmos genéticos paralelos, en especial bajo el modelo isla, para una plataforma GPU usando la tecnología CUDA.

Reconocimientos

Los autores agradecen el apoyo de la UNLPam y ANPCYT (proyecto 09F-049) y PICTO-UNLPAm (0278).

Referencias

1. Holland J., Adaptation in Natural and Artificial Systems. University of Michigan: Press, Ann Arbor 1975
2. Kirk D., Mei W., Hwu W., "Programming Massively Parallel Processors,ïn A Hands-on Approach., 2010, p. 280.
3. Lee V. et al, "Debunking the 100X GPU vs. CPU myth architecture ,ïn Proceedings of the 37th annual international symposium on Computer ISCA '10., 2010, p. 451.
4. NVIDIA, ÇUDA Toolkit 4.0"., http://developer.nvidia.com.
5. Karp R., Reducibility among combinatorial problems., 85103rd ed., Thatcher J (eds) Complexity of computer computations. In: Miller R, Ed., 1972.
6. Pardalos P., Resende M., Ribeiro C., Festa P., Randomized heuristics for the MAX-CUT problem. Optimization Methods and Software, 2002.
7. Rendl F., Helmberg C., A spectral bundle method for semidefinite programming, 2000.
8. Duarte A., Laguna M., Martí R., Advanced scatter search for the Max-Cut problem. , 2112638th ed., 2009.
9. Olivo O., Arráiz E., Competitive simulated annealing and tabu search algorithms for the Max-Cut problem. In: GECCO '09:proceedings of the 11th annual conference on genetic and evolutionary, computation. ACM, New York, pp 1797–1798, 2009.
10. Williamson D., Goemans M., Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J ACM 42(6):1115–1145, 1995.
11. Salto C., Alba E., Designing heterogeneous distributed AGs by efficiently self-adapting the migration period, DOI 10.1007/s10489-011-0297-9.
12. Lee V. et al., "Debunking the 100X GPU vs. CPU myth," in Proceedings of the 37th annual international symposium on Computer architecture - ISCA '10, 2010, p. 451.
13. Nannen, V. and Smit, S.K. and Eiben, A.E., "Costs and Benefits of Tuning Parameters of Evolutionary Algorithms," in Proceedings of the 10th international conference on Parallel Problem Solving from Nature: PPSN X, 2008, p. 528-538.

A novel Competitive Neural Classifier for Gesture Recognition with Small Training Sets

Facundo Quiroga, Leonardo Corbalán

III-LIDI Institute, Informatics Faculty, UNLP
La Plata, Buenos Aires, Argentina
fquiroga,corbalan@lidi.unlp.edu.ar

Abstract. Gesture recognition is a major area of interest in human-computer interaction. Recent advances in sensor technology and computer power has allowed us to perform real-time joint tracking with commodity hardware, but robust, adaptable, user-independent usable hand gesture classification remains an open problem. Since it is desirable that users can record their own gestures to expand their gesture vocabulary, a method that performs well on small training sets is required. We propose a novel competitive neural classifier (CNC) that recognizes arabic numbers hand gestures with a 98% success rate, even when trained with a small sample set (3 gestures per class). The approach uses the direction of movement between gesture sampling points as features and is time, scale and translation invariant. By using a technique borrowed from object and speaker recognition methods, it is also starting-point invariant, a new property we define for closed gestures. We found its performance to be on par with standard classifiers for temporal pattern recognition.

Keywords: gesture recognition,scale invariant,speed invariant,starting-point invariant,neural network,cpn,competitive

1 Introduction

The recent rise of gesture interaction as a practical possibility, through new devices and sensors, has made natural gesture-based software a reality, with applications ranging from web browsing and gaming to sign language interpretation and smart home interaction. A gesture recognition system usually consists of two stages: low-level feature extraction and representation based on sensor data, for example using depth images taken from a time-of-flight camera; and gesture classification employing the extracted features. Current research efforts in human-computer interaction, computer vision, motion analysis, machine learning and pattern recognition are contributing to the creation of even more robust and usable recognition systems [17] in both stages.

The Kinect SDK has been recently used as a stepping stone for doing research in the second stage of a gesture recognition system based on body joint 3D positions, for example to perform a comparison of template methods for real-time

gesture recognition [14], testing a Weighted Dynamic Time Warping algorithm [2], posture detection [15], to design a system that monitors posture ergnonomics [16], human behavior recognition by a mobile robot following human subjects [1], and allowing wheelchair-accessible gesture-based game interaction [5]. This is a recent trend in vision based gesture-recognition, since previous works mostly focused on the feature extraction stage [3], and used image based features instead of performing body joint tracking and using the resulting 3D position data to construct appropiate features as in [18] for hand-gesture recognition. Also, most hand gesture recognition research has additionaly employed finger and palm information because they tipically address sign language recognition.

In this work, we propose a speed, translation and scale-invariant method, the Competitive Neural Classifier (CNC), for recognizing hand gestures based on a time-labeled sequence of 3D hand positions, with a restricted training set size. The CNC was partially inspired by Probabilistic SOM's (ProbSOM) [4] approach to speaker recognition. The proposed methodology for the CNC (and ProbSOM) discards the sequence information contained in the extracted features computed from the sample data and thus follows an approach similar to those employed in object or speaker recognition, that is, the characterization of a sample by means of a set of distinctive features; an approach unexplored, to the best of our knowledge, in the area of gesture recognition. The architecture of each sub-classifier maps those features into a lower dimensional space by means of a competitive layer trained also in a competitive fashion, and uses the resulting activation patterns as a gesture signature employed as input for another competitive network that outputs the likelihood of each class given each sample. Finally, the use of many such sub-classifiers improves the recognition rate by combining different mappings derived from different clusterings and thus provides robustness to the method.

We focus directly on the the second stage of the gesture recognition by leveraging the Kinect SDK's recognition algorithms to obtain user joint positions and generate a gesture database to test the method and compare its performance against ProbSOM and other two known techniques: Input-Delay Feedforward Networks [7], and a modified Ordered Means Model [6] algorithm called ST-OMM.

This work is organized as follows: we introduce the gesture database in section 2, together with the preprocessing and feature extraction stage; then, we present the CNC in section 3, a brief introduction to the compared methods in section 4, and finish with experimental data and conclusions in sections 5 and 6.

2 Preprocessing and Feature Extraction

2.1 Gesture Database

We performed all of our experiments using the Arabic Numbers Hand Gesture Database [1], a small database of our creation with 10 samples of each of the arabic

[1] More information available at https://sites.google.com/site/dbanhg/

digits performed using the left hand and recorded with Microsoft Kinect's SDK, which gives a set of classes $C = 0 \dots 9$. The recording of all samples was done at an average of 28 *fps*, by the same person. In the recording of the different samples of each digit the orientation of the person with respect to the camera was the same, but the samples were performed starting from different hand and body positions and each gesture shape was drawn with different sizes.

Each gesture sample $\mathbf{s_i} \in S$, where S is our gesture database, consists of a sequence $\mathbf{s_i} = \mathbf{s_i}[1], \mathbf{s_i}[2], \dots, \mathbf{s_i}[n_i], \quad \mathbf{s_i}[j] \in \mathbb{R}^3, \quad j = 1 \dots n_i$, corresponding to the hand positions in a 3D space, time-labeled $T_i = t_1, \dots, t_{n_i}, \quad t_j \in \mathbb{R}, \quad 0 = t_1 < t_2 < \dots < t_{n_i}$, and gesture class label c_i . Each sample $\mathbf{s_i}$ may have a different number of positions n_i, depending on the capture frame rate and the time used to execute the gesture.

2.2 Preprocessing

In the preprocessing stage, the first and last 3 positions of each sample are discarded because they usually contain unwanted information introduced by incorrect gesture segmentation. The samples were smoothed individually using an unweighted moving average with a window of size w in order to remove the high-frequency information from the signal, because the chosen features are based on the direction between consecutive sample points and small fluctuations in direction give too local information to characterize the overall shape of a gesture.

The nature of the feedforward and ST-OMM architectures requires the number of sample positions n_i to be constant across all samples. Also, it is desirable to obtain speed-invariant features. In order to achieve this each sample is resampled to a sequence of constant length n using cubic splines interpolation with an arc-length parameterization.

The parameterization of each sample $\mathbf{s}$ of length q gives the position of the hand in the 3D space as a function of the arc-length distance from the first position of the sample. For each position $\mathbf{s}[j]$ we calculate the arc-length distance from the first position $l_j = \sum_{k=2}^{j} ||\mathbf{s}[j] - \mathbf{s}[j-1]||$, where $||\cdot||$ is the Euclidean norm. The resampling is done at n uniformly distributed knots in the total sample arc length $L = l_q$, given by $k_j = \frac{j-1}{n-1} * L, \quad j = 1 \dots n$. We obtain the sequence of points $\mathbf{r}$ of length n such that $\mathbf{r}[j] = cubic(k_j, near_4(k_j)) \quad j = 1 \dots n$ where $cubic(x, (x_1, x_2, x_3, x_4))$ performs the cubic interpolation at distance x using distances (x_1, x_2, x_3, x_4) whose positions are known and $near_4(x)$ returns the 4 distances nearest to the distance x (ie, $min_4 = (|l_1 - x|, \dots, |l_q - x|)$) such that $x_1 < x_2 \leq x \leq x_3 < x_4$.

2.3 Features

From the smoothed, resampled sequence $\mathbf{r_i}$ we compute the normalized first difference $\mathbf{d_i}$ where $\mathbf{d_i}[j] = \frac{\mathbf{r_i}[j+1] - \mathbf{r_i}[j]}{||\mathbf{r_i}[j+1] - \mathbf{r_i}[j]||}, \quad j = 1 \dots n-1, \quad \mathbf{d_i}[j] \in \mathbb{R}^3$, which represents the relative directions between the sample positions. By computing the first difference, we obtain a translation invariant representation. By normalizing, we remove all speed information contained in the length of each direction

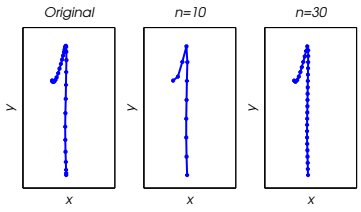


Fig. 1. Projection on the xy plane of a sample of the gesture 1, before and after resampling with $n = 10, 30$. The greater density of sampling points, shown as dots, in the upper part in relation to the lower part of the original that was caused by a difference in speed when performing the gesture has been compensated by resampling with different degrees of detail.

vector, making the signal invariant to the speed and scale with which the gesture was drawn. Note that without the resampling the normalizing would still leave a considerable amount of velocity information in the signal because the amount of sampling points in the segments (in the arc-length sense) where the user performs the gesture at high speed is lower than in those segments where the hand moves more slowly.

As an alternative to the direction vectors, we also employed the angles of the spherical representation of the direction vectors, obtaining a representation $\mathbf{a_i}$, where $\mathbf{a_i}[j] = (\theta, \phi)_j \in \mathbb{R}^2$. The z coordinate is left out because it is always 1 as a consequence of the previous normalization, and we rescaled the angles so that $\theta, \phi \in [-1, 1]$.

Given the periodical nature of the angle representation, in all angle differences calculated for all classification algorithms we utilized the difference function $d(\alpha, \beta) = min(|\alpha - \beta|, 2 - |\alpha - \beta|)$. Although both features are mathematically equivalent and share the same desirable properties (translation, scale and speed invariance [12]), they produce slightly different results in our experiments. In the following sections we will refer to a secuence of sample features as $\mathbf{s}$, where $\mathbf{s}[j] \in \mathbb{R}^d$ could represent either feature (with $d = 2$ for angles and $d = 3$ for cartesian directions).

2.4 Starting point invariance

We define the *starting-point invariance* property for closed gestures, which are those that, ideally, start and finish in the same position. In such a case, like when recognizing the gesture corresponding to the digit 0, we would like the recognition algorithm to be able to detect the gesture without regard to where the user started performing it. Therefore, a feature given by $f :: Samples \rightarrow Features$ is starting-point invariant if $f(\mathbf{s}) = f(shift(\mathbf{s}, k)), \quad k = 1..n-1$ where n is the length of the sample, $shift(\mathbf{s}, k) = (s_{(k)\%(n+1)}, s_{(k+1)\%(n+1)}, \ldots, s_{(n+k)\%(n+1)})$ and $\%$ is the *modulo* operator.

3 Competitive Neural Classifier (CNC)

The CNC, a new gesture recognition strategy based on competitive neural networks, employs a combination of competitive neural networks and a decision mechanism that integrates information generated by multiple subclassifiers to add robustness to the recognition process and improve the recognition rate. The structure of a CNC with p subclassifiers is shown in figure 2 a).

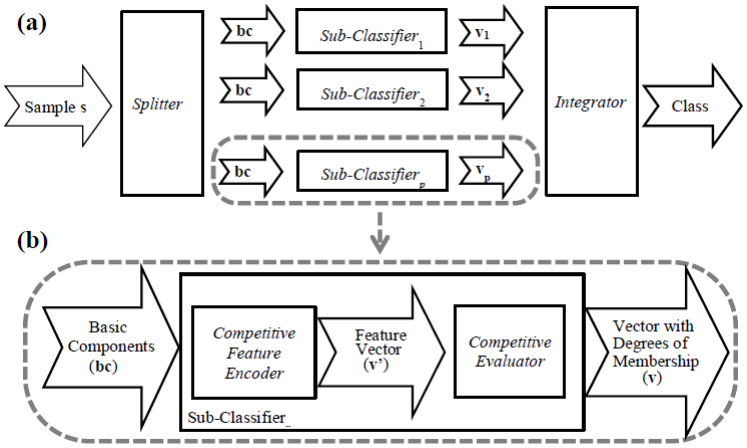


Fig. 2. a) CNC Architecture b) Subclassifier module

3.1 Classifier structure and functioning

When a sample **s** enters the classifier, the Splitter module generates the sequence of basic components **bc** and sends a copy to each of the p subclassifiers. Note that each basic component **bc**$[j]$ is simply the cartesian or angle representation of the jth direction vector described in the previous section. The Competitive Feature Encoder Module (CFEM) in each subclassifier (figure 2 b) implements a neural network nn' with m neurons $h'_1, h'_2, \ldots, h'_m$, trained with the well-known CPN learning algorithm [8]. Given a sequence of n basic components (a sample), the CFEM maps it into a characteristic vector $\mathbf{v}'$ according to:

$$\mathbf{v}' = (v'_1, v'_2, \ldots, v'_m)$$
$$v'_k = count(h'_k)/n \quad k = 1..m$$

where $count(h'_k)$ represents the number of basic components for which the neuron h'_k had an activation value greater than other neurons. Therefore, $\mathbf{v}'$ codes information about the distribution of each basic component of the sample **s**

according to the hidden space. Since the $nn's$ are trained independently, each will produce a classification based on different clusterings, hopefully complementary to each other.

The Competitive Evaluator Module (CEM) contains a neural network nn with a competitive layer composed of z hidden neurons $h_1, \ldots, h_z$ with corresponding weights $\mathbf{w_1}, \ldots, \mathbf{w_z}$, where z is determined by the learning algorithm. The neurons of this layer are stimulated with the vector $\mathbf{v'}$ to output a vector $\mathbf{v} = (v_1, \ldots, v_z)$ of scores. The network deviates from the typical competitive architecture in that instead of identifying a unique winner neuron, this vector represents the degree of membership of the sample to each neuron using the inverse of the Manhattan distance $||\cdot||$ as a similarity function, and is given by $v_k = 1/||(\mathbf{v'}, \mathbf{w_k})|| k = 1..z$.

Each neuron h_k is associated with a gesture class c by means of a function $f :: Neuron- > Class$, and therefore such scores $\mathbf{v}$ also represent the degree of membership of the sample to each class. Finally, the Integrator module receives the outputs $\mathbf{v_i}, \quad i = 1..p$ of every classifier, and calculates the corresponding class as $class = f(max_k(scores))$, where $scores = \sum_{i=1}^{p} \mathbf{v_i}$ and max_k returns the index of the vector component (hidden neuron) with the maximum value.

3.2 Learning algorithm

The subclassifiers are trained independently in two stages. First, the nn' of each CFEM is trained with the classical CPN iterative learning algorithm using the basic components of all samples and the Manhattan distance as a similarity function.

After the CFEM's training is finished, each h'_i corresponds to the centroid of a cluster of basic components. Given the nontraditional use of each nn' in the generation of characteristic vectors $\mathbf{v'}$on the training algorithm can be stopped very quickly (as early as two iterations in our tests) while still obtaining good results.

The neural network nn of the CEM requires no training. Once the CFEM's training is complete, the nn is built with $z = |C| \times u$ neurons h_i, where u is the number of samples of each class and $|C|$ the number of classes. To each h_i corresponds a weight vector $\mathbf{w_i} = \mathbf{v'_i}$ where $\mathbf{v'_i}$ is the characteristic vector generated by the CFEM when presented with the basic components of sample $\mathbf{s_i}$. The mapping function f is simply $f(i) = c_i$, that is, given sample index i, it returns the class label of sample i.

4 Compared Methods

4.1 Probabilistic SOM (ProbSOM)

ProbSOM combines the well know Self Organizing Map (SOM) clustering technique with a probabilistic decision model which has been succesfully applied to solve speaker identification problems. CNC has been partially inspired by

ProbSOM's approach of extracting basic components of a signal to stimulate a competitive layer, but differs substantially in the mapping approach and decision process. For an accurate comparison with CNC, we tested an ensemble of p independently trained ProbSOMs used as subclassifiers, and use as final output the mean of the subclassifiers scores for each class.

4.2 Input Delay Feedforward Network (ID-FF)

Input Delay Feedforward Networks (ID-FF) are a class of feedforward networks adapted for temporal sequence processing. Applied to this type of problems, their main distinctive feature is the grouping of the sequence of feature vectors that characterize a gesture into a single vector by simply aggregating all the features while maintaining their temporal ordering. This results in a representation that simply but effectively models the dynamics of the gesture and has the advantage that the neural network can be designed and trained in a traditional way.

We chose the standard 2-layered feedforward network, trained with resilient backpropagation, a well-known, fast and proven first-order training algorithm [9]. We employed the transfer function *tansig* for both the hidden and output layer. The output layer contains 10 neurons, one for each gesture class c, with output values $o_c \in -1, 1, c = 1 \dots 10$.

4.3 Small-Training-Set Ordered Means Model (ST-OMM)

The Ordered Means Model (OMM) is a simplified Hidden Markov Model (HMM) which has been successfully applied for solving gesture recognition problems in a variety of settings. Moreover, HMMs are the de-facto standard for generative classifier models, (and, arguably, gesture recognition [10]) and thus a good choice for comparison.

While most model building methods for both the OMM and HMM commonly employ some variant of the Expectation-Maximization (EM) algorithm to find optimal values for their parameters, such a choice is ill-suited for small training sets as required for our problem. Therefore, we have created the Small-Training-Set-OMM (ST-OMM), an adaptation of the OMM approach to comply with this requirement.

The ST-OMM takes as input a sample, which is a fixed length sequence $\mathbf{s}$ of n features, and outputs the likelihoods of the sample belonging to each gesture class c. The ST-OMM is built with n competitive Gaussian Mixture Models (C-GMM), and each C-GMM $G_j, j = 1 \dots n$ is composed of a set of states $S_{j,c}, c \in C$, with constant mixture coefficients $\omega_c = 1/|C|$. To every state $S_{j,c}$ corresponds to a set of Gaussian pdfs with means $\mu_{j,c,k}$ and covariance matrices $\mathbf{\Sigma}_{j,c,k}, k = 1 \dots |Sc|$, where Sc is the set of samples of class c, which models the probability of an emission of gesture part j by the gesture class c in a competitive fashion. This parameters are estimated as $\mu_{j,c,k} = \mathbf{s}_{ck}[j]$ and $\mathbf{\Sigma}_{j,c,k} = cov(I_{j,c})$, $k = 1 \dots |Sc|$, where $\mathbf{s}_{ck}$ is the kth sample of class c and $I_{j,c}$ is a matrix whose columns are $\left[\, \mathbf{s}_{c1}[j] \; \mathbf{s}_{c2}[j] \; \cdots \; \mathbf{s}_{c|Sc|}[j] \;\; {\scriptstyle k=1 \dots |Sc|} \,\right]$, that is, $I_{j,c}$ is the matrix that contains the j^{th} feature of every sample of class c.

We are therefore using each part of every sample of the training set - the best likelihood estimators for that set - as a C-GMM mean, which does not yield a computationally demanding model because we are specifically targeting a very small training sets. For the classification of a new sample $\mathbf{s}$, we calculate the likelihood of emission for each state $S_{j,c}$ as:

$$P(S_{j,c}|s[j]) = \frac{P(s[j]|S_{j,c})P(S_{j,c})}{P(s[j])} = \frac{p_{j,c}P(S_{j,c})}{\sum_{k\in C} p_{j,k}P(S_{j,k})} = \frac{p_{j,c}}{\sum_{k\in C} p_{j,k}}$$

where $P(S_{j,k}) = \frac{1}{|C|}$ is the same for all classes k, and $p_{j,k} = P(s[j]|S_{j,c}) = max(\mathcal{N}(s[j]; \mu_{j,c,k}, \mathbf{\Sigma}_{j,c,k}))$ is the maximum of the scores that model the likelihood of the j^{th} feature of the sample belonging to class c.

Then, the likelihood of the whole sample belonging to class c is:

$$P(c|s) = \frac{P(s|c)P(c)}{\sum_{k\in C} P(s|k)P(k)} = \frac{P(s|c)}{\sum_{k\in C} P(s|k)}$$

where we define $P(k) = \frac{1}{|C|}$ and $P(s|k) = \frac{\sum_{j=1}^{n} P(S_{j,k}|s[j])}{n}$

We can thus picture the ST-OMM as a $|C| \times n$ state HMM, with one left-to-right submodel for each gesture that does not allow a transition from a state to itself, that is:

$$P(S_{j,c} \to S_{j',c}) = \begin{cases} 1 & \text{if } j' = j+1 \\ 0 & \text{otherwise} \end{cases}$$

This restriction avoids doing a dynamic programming search over state combinations and, although it is obviously of lower computational capacity than a full HMM, works well given a large enough n, since any desynchronization between a novel performance of the gesture and the model produces only small mismatches.

5 Experimentation

We compared the recognition rate of the four methods on the gesture database, using the same preprocessing parameters for all. The resampling size n was set to 60 and the smoothing window size w to 5. Each algorithm was tested 500 times using random subsampling cross-validation, with 3 samples per class for the training set (30 total) and 7 for the test set (70 total). In the case of the feedforward network, 2 samples of each class were taken from the test set to be used as the validation set, leaving 5 for the test set. For the ID-FF, CNC subclassifiers and ProbSOM networks, we show results for hidden neurons $m = 30$, $m = 70$ and $m = 100$ respectively, which gave the best results in our experiments (we tested $m \in 5, 10, \ldots, 150$ to determine those values). Both the CNC and ProbSOM had $p = 5$ sub-classifiers. In all experiments, for ProbSOM, ST-OMM

and ID-FF the class assigned to a sample gesture was the one that gave the best score for that class, irrespective of its actual value; that is, when presented with a sample, the method calculates the outputs o_c for each class normally and the corresponding class is assigned according to the rule $class = max_c(o_c)$.

Method / Feature	Direction	Angles
CNC	98.91 (1.20)	95.32 (2.20)
ProbSOM	62.69 (5.94)	43.57 (6.86)
ID-FF	91.54 (9.06)	80.88 (8.81)
ST-OMM	95.54 (2.72)	96.40 (2.09)

Table 1. Sample mean and standard deviation of recognition rates over 500 experiments, with training and test cases chosen randomly, $w = 5$, $n = 60$.

We also show the performance without resampling for the ProbSOM and CNC which do not require a fixed length input vector, although in this case they are not truly speed independent.

Method / Feature	Direction	Angles
CNC	98.37 (1.63)	83.30 (4.40)
ProbSOM	56.83 (4.67)	39.57 (6.86)

Table 2. Mean and standard deviation of recognition rates without resampling over 500 trials, $w = 5$.

As we can see, the recognition rates are slightly lower but not significantly, which shows that in principle the method could be used without a proper normalization in cases where the need for reduced computation outweighs that of high recognition rates.

6 Conclusion

A novel approach for gesture recognition with small training samples has been presented that is time, scale and translation invariant, and for closed gestures, starting-point invariant as well, while achieving high recognition rates comparable to other approaches and with a learning algorithm that requires few iterations to converge.

In our experiments, the CNC, ST-OMM and the ID-FF all perform reasonably well, especially first two. In addition, the CNC has the advantage of being quick to train and starting-point invariant without any modification, although at a greater intrinsic computational cost. For the other two methods, the same effect can be achieved by shifting the sample in a circular way in all n possible combinations which also scales up by a factor n their execution order.

In future work, we hope to determine the method's performance on larger gesture databases, and apply it in a real-time setting to test its ability to recognize gestures in a unsegmented stream of hand positions. Finally, we intend to improve and extend our current database with 3D gesture data to provide a reference point for future comparisons and benchmarkings.

References

1. Human behavior recognition by a mobile robot following human subjects. In: Evaluating AAL Systems Through Competitive Benchmarking (2013)
2. Celebi, S., Aydin, A.S., Temiz, T.T., Arici, T.: Gesture recognition using skeleton data with weighted dynamic time warping. Computer Vision Theory and Applications. Visapp (2013)
3. Chaudhary, A., Raheja, J.L., Das, K., Raheja, S.: Intelligent approaches to interact with machines using hand gesture recognition in natural way: a survey. arXiv preprint arXiv:1303.2292 (2013)
4. Estrebou C., Lanzarini L., H.W.: Voice recognition based on probabilistic som. In: Latinamerican Informatics Conference. CLEI 2010. Paraguay. October 2010. (2010)
5. Gerling, K.M., Kalyn, M.R., Mandryk, R.L.: Kinectwheels: wheelchair-accessible motion-based game interaction. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems. CHI EA '13, ACM (2013)
6. Grossekathofer, U., Wohler, N.C., Hermann, T., Kopp, S.: On-the-fly behavior coordination for interactive virtual agents: a model for learning, recognizing and reproducing hand-arm gestures online. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3. AAMAS '12 (2012)
7. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edn. (1998)
8. Hecht-Nielsen, R.: Counterpropagation networks. Applied optics 26(23), 4979–4983 (1987)
9. Igel, C., Husken, M.: Empirical evaluation of the improved rprop learning algorithms (2003)
10. Just, A., Marcel, S.: A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition. Comput. Vis. Image Underst. 113(4), 532–543 (Apr 2009), http://dx.doi.org/10.1016/j.cviu.2008.12.001
11. Kindratenko, V.V.: On using functions to describe the shape. J. Math. Imaging Vis. 18(3) (May 2003)
12. Lai, K., Konrad, J., Ishwar, P.: A gesture-driven computer interface using kinect. In: Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on. pp. 185–188 (2012)
13. Le, T.L., Nguyen, M.Q., Nguyen, T.T.M.: Human posture recognition using human skeleton provided by kinect. In: Computing, Management and Telecommunications (ComManTel), 2013 International Conference on (2013)
14. Martin, C., Burkert, D., Choi, K., Wieczorek, N., McGregor, P., Herrmann, R., Beling, P.: A real-time ergonomic monitoring system using the microsoft kinect. In: Systems and Information Design Symposium (SIEDS), 2012 IEEE (2012)
15. Sarkar, A.R., Sanyal, G., Majumder, S.: Article: Hand gesture recognition systems: A survey. International Journal of Computer Applications (2013)
16. Shen, X., Hua, G., Williams, L., Wu, Y.: Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields. Image and Vision Computing 30(3), 227 – 235 (2012), http://www.sciencedirect.com/science/article/pii/S0262885611001193, <ce:title>Best of Automatic Face and Gesture Recognition 2011</ce:title>

XIII WORKSHOP PROCESAMIENTO DISTRIBUIDO Y PARALELO

- WPDP -

CACIC2013

XIII WORKSHOP PROCESAMIENTO DISTRIBUIDO Y PARALELO - WPDP -

ID	Trabajo	Autores
5877	**Análisis de la escalabilidad y el consumo energético en soluciones paralelas sobre cluster de multicores y GPU para un problema con alta demanda computacional**	Erica Montes de Oca (UNLP), Laura De Giusti (UNLP), Armando E. De Giusti (UNLP), Marcelo Naiouf (UNLP)
5682	**N-Body Simulation Using GP-GPU: Evaluating Host/Device Memory Transference Overhead**	Sergio Martín (UNLaM), Fernando Tinetti (UNLP), Nicanor Casas (UNLaM), Graciela De Luca (UNLaM), Daniel Giulianelli (UNLaM)
5684	**Procesamiento de Señales SAR: Algoritmo RDA para GPGPU**	Mónica Denham (UNRN), Javier Areta (UNRN), Isidoro Vaquila (INVAP), Fernando Tinetti (UNLP)
5837	**Parallel implementation of a Cellular Automata in a hybrid CPU/GPU environment**	Emmanuel N. Millán (UNCUYO), Paula Martínez (UNCUYO), Verónica Gil Costa (UNSL), María Fabiana Piccoli (UNSL), Marcela Printista (UNSL), Eduardo M Bringa (UNCUYO), Carlos Garcia Garino (UNCUYO), Carlos Bederian (CONICET)
5893	**Evaluating tradeoff between recall and performance of GPU Permutation Index**	Mariela Lopresti (UNSL), Natalia Miranda (UNSL), Mercedes Barrionuevo (UNSL), María Fabiana Piccoli (UNSL), Nora Reyes (UNSL)
5785	**Managing Receiver-Based Message Logging Overheads in Parallel Applications**	Hugo Meyer (UAB), Dolores Isabel Rexachs del Rosario (UAB), Emilio Luque Fadón (UAB)
5621	**Optimizing Multi-Core Algorithms for Pattern Search**	Verónica Gil Costa (UNSL), Cesar Ochoa (UNSL), Marcela Printista (UNSL)
5797	**A tool for detecting transient faults in execution of parallel scientific applications on multicore clusters**	Diego Montezanti (UNLP), Enzo Rucci (UNLP), Dolores Isabel Rexachs del Rosario (UAB), Emilio Luque (UAB), Marcelo Naiouf (UNLP), Armando E. De Giusti (UNLP)

XIII WORKSHOP PROCESAMIENTO DISTRIBUIDO Y PARALELO - WPDP -

ID	Trabajo	Autores
5676	**Lessons learned from contrasting a BLAS kernel implementations**	Andrés More (INTEL)
5880	**Mejoras en la eficiencia mediante Hardware Locality en la simulación distribuida de modelos orientados al individuo**	Silvana Lis Gallo (UNLP), Francisco Borges (UAB), Remo Suppi (UAB), Emilio Luque (UAB), Laura De Giusti (UNLP), Marcelo Naiouf (UNLP)
5892	**Un método de optimización para mejorar la salida de un modelo computacional de cuenca de ríos**	Adriana Angélica Gaudiani (UNGS), Emilio Luque Fadón (UAB), Armando E. De Giusti (UNLP), Marcelo Naiouf (UNLP)
5894	**Multithreading model for evacuations simulation in emergency situations**	Pablo Cristian Tissera (UNSL), Marcela Printista (UNSL), Emilio Luque (UAB)
5895	**Efficiency analysis of a physical problem: Different parallel computational approaches for a dynamical integrator evolution**	Adriana Angélica Gaudiani (UNGS), Alejandro Soba (CNEA), María Florencia Carusela (UNGS)
5725	**Desarrollo de aplicaciones paralelas en Erlang/ OTP utilizando múltiples planificadores**	Juan Pisani (UNSL), Pablo Cristian Tissera (UNSL), Marcela Printista (UNSL)
5768	**Parameters Calibration for Parallel Differential Evolution based on Islands**	Laura Tardivo (UNRC), Paola Caymes Scutari (UTN-FRM), Miguel Mendez Garabetti (UTN-FRP), Germán Bianchini (UTN-FRM)
5769	**Cómputo en Paralelo para Integrales Multicéntricas usando una Distribución Balanceada**	Ana Rosso (UNRC), Claudia Denner (UNRC), Guillermo Fraschetti (UNRC), Laura Tardivo (UNRC), Jorge Perez (UNRC), Juan Cesco (UNSL)

XIII WORKSHOP PROCESAMIENTO DISTRIBUIDO Y PARALELO - WPDP -

ID	Trabajo	Autores
5790	**Predicting the communication pattern evolution for scalability analysis**	Javier Panadero (UAB), Alvaro Wong (UAB), Dolores Isabel Rexachs del Rosario (UAB), Emilio Luque (UAB)
5783	**Arquitectura en capas para acceso remoto SAD**	Karina M. Cenci (UNS), Leonardo De Matteis (UNS), Jorge Ardenghi (UNS)

Análisis de la escalabilidad y el consumo energético en soluciones paralelas sobre cluster de multicores y GPU para un problema con alta demanda computacional

Erica Montes de Oca[1], Laura De Giusti[1], Armando De Giusti[1 2], Marcelo Naiouf[1].

[1] Instituto de Investigación en Informática LIDI (III-LIDI)
Facultad de Informática, Universidad Nacional de La Plata.
La Plata, Buenos Aires, Argentina.
[2] CONICET
{emontesdeoca,ldgiusti,degiusti,mnaiouf}@lidi.info.unlp.edu.ar

Resumen. Este trabajo realiza un estudio de la escalabilidad y el consumo energético, en el uso de un cluster de multicore y una placa de GPU con 384 cores, teniendo como caso de aplicación el problema de los N-body. Se implementaron una solución paralela en memoria compartida para CPU usando Pthread, una solución en memoria compartida para GPU usando CUDA y una solución en memoria distribuida en CPU utilizando MPI. Se presentan y analizan los resultados obtenidos, que muestran en este problema que el uso de la GPU no solo logra acelerar el cómputo sino también, reducir el consumo energético.

Palabras claves: multicore, cluster de multicores, GPU, N-body, escalabilidad, Green Computing, consumo energético.

1. Introducción

La tecnología ha permitido mejorar la calidad de vida mundial. Los avances en los procesadores en los últimos tiempos, han logrado acelerar la solución muchos problemas de la vida real. La llegada de los multicores, no sólo ha permitido disminuir los tiempos de cómputo de varias aplicaciones [1], sino que también ha logrado reducir el consumo energético de los procesadores; ya que los mismos están formados por más procesadores pero mucho más simples.

Los multicores pueden agruparse formando clusters en los que se combina la memoria compartida por los núcleos de un multicore con la memoria distribuida para la comunicación entre multicore [2], dando lugar a un esquema híbrido. Por otra parte, en las últimas décadas, una nueva plataforma de propósito específico se ha abierto paso como una alternativa para el Cómputo de Altas Prestaciones: la GPU [3] [4] [5] [6].

Sin embargo, la aceleración del cómputo trae aparejada la problemática de la gran cantidad de energía consumida, que se ha convertido en un aspecto significativo tanto en la fabricación del hardware como en las implementaciones software.

Desde una perspectiva actual, es nuestra responsabilidad proveer no solo un avance tecnológico sino un uso responsable del mismo [7]. Ya sea desde el hardware o el software, debe pensarse en reducir gastos innecesarios en el consumo energético. En

ese sentido, este trabajo presenta una comparación de soluciones al problema de la atracción gravitacional de los cuerpos celestes utilizando cluster de multicores y GPU.

El trabajo está organizado de la siguiente manera: en la Sección 2 se introduce al concepto de Green Computing; la Sección 3 plantea un breve comentario del problema de los N-body, mientras que en la Sección 4 se muestran los resultados experimentales obtenidos. La Sección 5 presenta las conclusiones y trabajos futuros.

2. Green Computing

Green Computing comprende el uso eficiente de los recursos. Su objetivo es reducir el uso de materiales peligrosos, maximizar la eficiencia de la energía durante el ciclo de vida de producción, y promover el reciclado o biodegrabilidad de los productos y desechos fabriles [8].

En muchos casos los consumidores no toman en cuenta el impacto ecológico a la hora de comprar sus computadoras, solo prestan atención a la velocidad de sus prestaciones y al precio. Sin embargo, a mayor velocidad de procesamiento se requiere mayor poder energético, trayendo el problema de la disipación del calor, que necesita más energía eléctrica para mantener al procesador en la temperatura normal de trabajo. Los diseñadores de hardware ya han planeado varias estrategias para colaborar con la reducción del consumo energético, desde la fabricación del equipo hasta su reciclado [9].

Un gran avance se ha realizado con el paso del monoprocesador a los procesadores multicore, ya que estos últimos suelen ser más simples, por lo que hacen un uso más eficiente de la energía. Las grandes compañías (Intel y AMD), han tomado conciencia sobre esta necesidad del uso eficiente de los recursos en la producción de los mismos [10] [11].

En los últimos tiempos, la eficiencia energética se ha transformado en uno de los factores influyentes en el desarrollo de aplicaciones. En el campo del Computación de Altas Prestaciones (HPC, High Performance Computing), se están realizando investigaciones orientadas no solo a disminuir la energía consumida, sino también en un sistema de gestión de energía que dada una aplicación y un plataforma HPC presente alternativas de ejecución dependientes de: el consumo energético, la potencia máxima (capacidad de la infraestructura eléctrica y el equipo de refrigeración) y el rendimiento [12]. En HPC la eficiencia energética es un factor que limita el desarrollo de aplicaciones, ya que la cantidad de energía necesaria para procesar las grandes cantidades de datos se ha convertido en un problema al cual cada vez se debe prestar más atención.

Además de la eficiencia energética, se debe tener en cuenta: la escalabilidad de los sistemas y el precio de la energía eléctrica. La gran cantidad de energía consumida reduce la escalabilidad de los sistemas, lo que los hace menos útiles a largo plazo. Por lo tanto, no solo se debe estudiar y analizar la escalabilidad con respecto al problema y la arquitectura, sino también se debe orientar la misma al consumo total final de las aplicaciones, para que por un lado no supere la cantidad de energía eléctrica que se nos suministra y por el otro, esté acorde a los gastos económicos disponibles.

3. Caso de estudio: problema de los N-body

El problema de los N-body es clásico en el cómputo científico y ha sido muy estudiado por su adaptabilidad a distintas aplicaciones del mundo real [13]. El presente trabajo está centrado en la aplicación de la fuerza de atracción gravitacional, basada en la teoría de Newton sobre su Ley de Atracción, que enuncia: "La fuerza de gravedad entre dos cuerpos es proporcional a sus masas e inversamente proporcional al cuadrado de sus distancias" [14].

El modelado del problema requiere del conocimiento de la siguiente información: la masa, la velocidad, la posición y la fuerza de atracción inicial de cada cuerpo. La Ecuación (0) es el cálculo central de todo el procesamiento, y se basa en la fuerza de magnitud de cada cuerpo [15].

$$F = (G \times m_i \times m_j) / r^2 \qquad \textbf{(0)}$$

con r = distancia, G = cte gravitacional $(6{,}67 \times 10^{11})$

El algoritmo secuencial en el que se basan las soluciones paralelas es denominado *all pair* o *simulación directa*. Todos los cuerpos que conforman el espacio del problema calculan su fuerza de atracción gravitacional con el resto. Por lo que la complejidad del problema es de $O(N^2)$ [16]. Para optimizar el acceso a la memoria caché y reducir el tiempo de ejecución, se realiza un procesamiento de los datos por bloques, siendo el tamaño de cada bloque el óptimo para la arquitectura empleada. Los algoritmos paralelos implementados para memoria compartida (en Pthread) y memoria distribuida (en MPI), también hacen uso del procesamiento por bloques para reducir los tiempos de ejecución al disminuir los fallos de caché.

En el caso del algoritmo de memoria compartida, un thread principal es el encargado de realizar la inicialización de los datos, y luego repartir a cada uno de los T-1 threads que conforman la simulación, bloques de datos del tamaño óptimo de la caché. Para la resolución del problema, se crean T-1 threads especificados por el usuario, ya que el hilo principal también realiza trabajo. Una vez que todos los threads tienen delimitado los datos a procesar, comienza el cálculo de la fuerza de atracción gravitacional para cada cuerpo de su conjunto de datos. Este cálculo se repetirá por cada paso de simulación, teniendo que esperar cada thread en una barrera de sincronización antes de comenzar un nuevo paso, para que todos los demás hilos puedan disponer de los datos actualizados del paso anterior.

Para la solución desarrollada con MPI, la implementación del algoritmo es similar con la excepción de que se trata de procesos y no de threads, y que una vez que el proceso terminó de realizar los cálculos para dicho paso, deberá comunicar los resultados obtenidos, a los demás procesos, antes de comenzar un nuevo paso de simulación.

En el algoritmo desarrollado para la GPU en CUDA, la inicialización de los datos se realiza en la CPU, y luego los mismos son comunicados por medio de PCI-E a la GPU. Una vez que la GPU tiene los datos copiados en su memoria global, cada thread que conforma la ejecución copiará los datos correspondientes para realizar los cálculos de la fuerza de atracción gravitacional a la memoria shared, para que de esta forma se optimice el acceso a la memoria. Como la memoria compartida o shared

tiene un tamaño reducido [17] [18] [4] [19], la copia de los datos desde la memoria global a la misma se hace por bloques del mismo tamaño del bloque de thread. Dicha transferencia de datos entre las memorias se lleva a cabo tantas veces hasta que todos los datos necesarios hayan sido copiados para realizar los cálculos correspondientes. Una vez realizado el procesamiento, la GPU enviará por medio de PCI-E los resultados a la CPU.

En [20] se describe con mayor detalle el resto de las ecuaciones utilizadas en las soluciones, así como una exposición más minuciosa de las soluciones paralelas implementadas.

4. Resultados experimentales

El entorno de prueba está compuesto por las siguientes arquitecturas:

- Un Cluster de Multicore con procesadores Quad Core Intel i5-2300 de 2,8 GHz, con caché de 6MB, y Sistema Operativo Debian de 64 bits.
- Una GPU Geforce TX 560TI, con 384 procesadores con una cantidad máxima de threads de 768 y 1 GB de memoria RAM.

La cantidad de cuerpos para cada simulación es variada entre 65535, 128000 y 256000 cuerpos para dos pasos de la simulación. Los datos obtenidos son el resultado de un promedio de varias ejecuciones.

Para la solución en memoria distribuida se realizó la ejecución utilizando una máquina por proceso (por lo que la comunicación de los mismos se realiza por medio de la red). Además, se llevó a cabo una ejecución de todos los procesos en una sola máquina.

En la Tabla 1 se muestran los tiempos de ejecución en segundos para dos pasos de simulación del problema, de las soluciones en CPU, tanto en memoria compartida como en memoria distribuida, utilizando 2 y 4 procesadores. En la Tabla 2 se presentan los tiempos de ejecución (en segundos) obtenidos para la solución en GPU, empleando CUDA, con un tamaño de bloque de threads de 256, que es el tamaño óptimo para la arquitectura utilizada, y dos pasos de simulación.

Tabla 1. Tiempos de ejecución en segundos para los algoritmos de MPI (para las dos formas de ejecución) y Pthreads en CPU. Con P = cantidad de procesos y T = cantidad de threads.

Cantidad de cuerpos de la simulación	Pthread (T = 2)	MPI (P = 2)	MPI en una maq. (P = 2)	Pthread (T = 4)	MPI (P = 4)	MPI en una maq. (P = 4)
65535	101,68	96,63	92,11	53,87	50,56	53,91
128000	397,39	352,39	352,93	213,94	175,66	182,08
256000	1572,81	1417,61	1427,42	810,17	717,80	728,45

Tabla 2. Tiempos de ejecución en segundos para el algoritmo de CUDA en GPU. Con T = cantidad de threads por bloque.

Cantidad de cuerpos de la simulación	CUDA (T = 256)
65535	1,04
128000	3,97
256000	15,75

Para la medición del consumo energético se empleó un osciloscopio digital, con una resolución de 8 bits con dos entradas, una para capturar la información de la tensión y la otra para la corriente. Ésta última proviene de una pinza trasductora con sensibilidad de ajuste a los siguientes valores: 1A/100mV, 1A/10mV y 1A/1mV.

La tensión se midió directamente de la línea eléctrica a la cual se encuentra conectado el cluster de multicores. La información recogida por el osciloscopio digital es enviada a otro equipo para ser analizada. La información de la corriente es obtenida del cable de entrada de las fuentes de energía de la arquitectura utilizada.

El osciloscopio digital coloca los datos calculados en buffers de 10240 muestras (10 KB). Cada buffer representa un tiempo aproximado de 40 milisegundos, lo que da un intervalo de muestreo de 40 ms / 10 KB = 3,9 μs.

En la Tabla 3 se muestran los resultados medidos en Joules totales consumidos por la aplicación para las distintas configuraciones y soluciones en la arquitectura CPU, para dos pasos de simulación con un tamaño de problema de 65535, 128000 y 256000 cuerpos.

Tabla 3. Joules totales consumidos por los algoritmos de MPI y Pthread en CPU. Con P = cantidad de procesos y T = cantidad de threads.

Cantidad de cuerpos de la simulación	Pthread (T = 2)	MPI (P = 2)	MPI en una maq. (P = 2)	Pthread (T = 4)	MPI (P = 4)	MPI en una maq. (P = 4)
65535	740,74	660,05	699,72	497,14	970,46	428,23
128000	2858,60	2415,74	2840,93	1953,98	3237,15	1798,35
256000	11290,59	9235,81	10040,86	7971,86	13123,58	6394,27

La Tabla 4 presenta el consumo total en Joules del algoritmo en GPU para 65535, 128000 y 256000 cuerpos en dos pasos de la simulación. Las mediciones de consumo energético para la arquitectura GPU se realizaron de la misma forma que para la arquitectura en CPU. La información de consumo obtenida y analizada es resultado de medir en el cable de entrada de la fuente del equipo que contiene la placa GPU.

Tabla 4. Joules totales consumidos por el algoritmo de CUDA en GPU. Con T = cantidad de threads por bloque.

Cantidad de cuerpos de la simulación	CUDA (T = 256)
65535	13,67
128000	60,94
256000	239,12

Las Figuras 1 y 2 muestran el consumo total obtenido para los algoritmos en CPU de memoria compartida y distribuida, para los distintos tamaños del problema (65535, 128000 y 256000) en dos pasos de simulación, con una configuración de 2 y 4 procesadores respectivamente. En la Figura 3, se presenta el consumo total resultante de ejecutar la aplicación en GPU, utilizando bloques de 256 threads para 65535, 128000 256000 cuerpos en dos pasos de simulación.

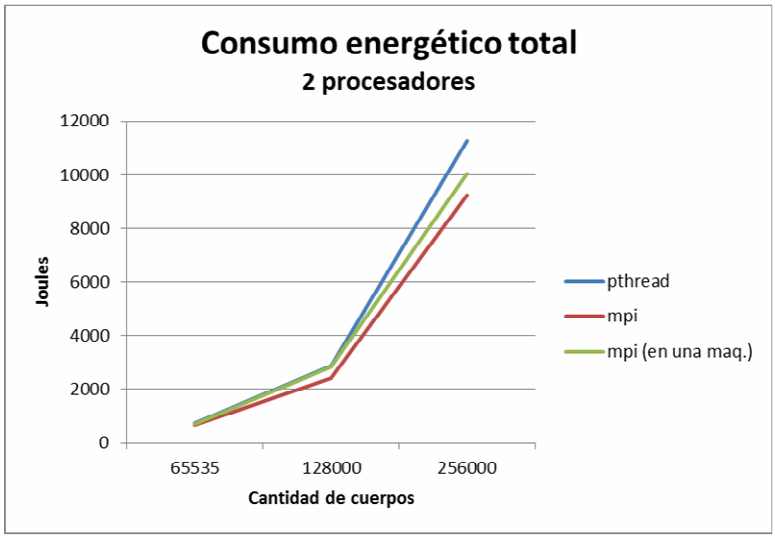


Fig. 1. Consumo total en Joules de los algoritmos MPI y Pthread con dos procesadores, para 65535, 128000 y 256000 cuerpos para dos pasos de simulación.

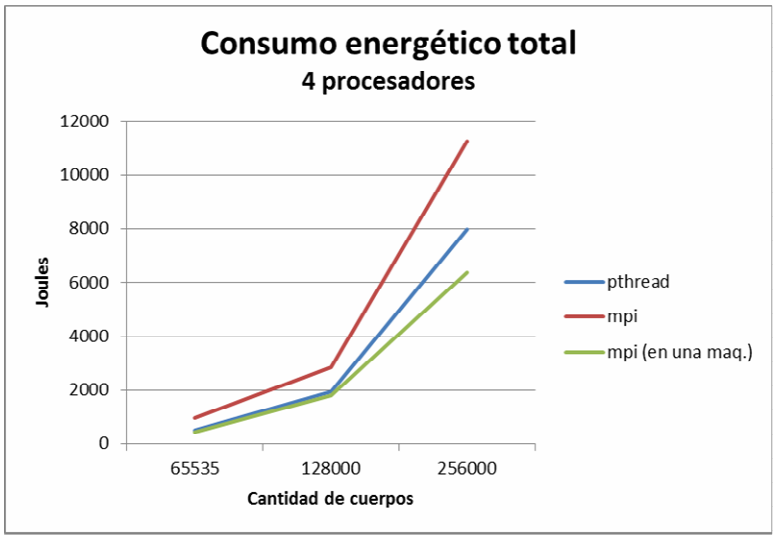


Fig. 2. Consumo total en Joules de los algoritmos MPI y Pthread con cuatro procesadores, para 65535, 128000 y 256000 cuerpos para dos pasos de simulación.

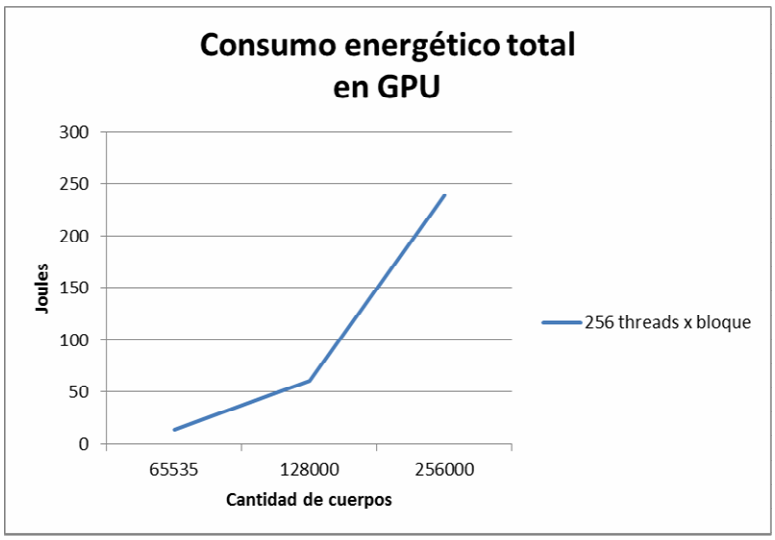


Fig. 3. Consumo total en Joules del algoritmo en CUDA para GPU con un tamaño de bloque de 256 thread, para 65535, 128000 y 256000 cuerpos para dos pasos de simulación.

A partir del entorno de prueba planteado, se obtuvieron los resultados anteriormente presentados. Se puede apreciar, que en cuanto a los tiempos de ejecución la diferencia entre las soluciones paralelas en CPU es poco importante. El tiempo de ejecución del algoritmo desarrollado en MPI para las ejecuciones con todos

los procesos en una máquina, y un proceso por máquina, es muy próximo entre sí. En cuanto a la GPU, notablemente se aprecia que el tiempo de ejecución obtenido para los distintos tamaños del problema es significativamente menor al de las soluciones paralelas desarrolladas para la arquitectura CPU.

Desde el punto de vista del consumo energético total se puede apreciar que el de la solución paralela en GPU es la que presenta un menor consumo, muy notable por la gran aceleración del cómputo obtenida. Mientras, que el consumo mayor alcanzado por la GPU se logra en el tamaño máximo del problema para las pruebas realizadas, el consumo total para el caso de las versiones paralelas en Pthread y MPI en una sola máquina, doblan dicho consumo para el tamaño inferior probado.

Para los casos de las soluciones paralelas en CPU, el aumento de procesadores aproxima el consumo energético total de las implementaciones en Pthread y MPI ejecutando todos los procesos en una sola máquina. Sin embargo, para el caso en el que cada proceso se ejecuta en una máquina distinta, el consumo se dispara al aumentar los procesadores, por razón de que se trata del consumo energético total de todas las máquinas completas.

La escalabilidad de un sistema paralelo refleja la capacidad del mismo de incrementar los recursos de procesamiento efectivamente. Como se puede observar, en los tiempos de ejecución obtenidos, el incremento de procesadores reduce el tiempo de procesamiento de la aplicación para los distintos tamaños del problema. Analizando a su vez, la escalabilidad desde el punto de vista del consumo energético total, para los casos de Pthread y MPI en una sola máquina el aumento de la arquitectura reduce el consumo por reducir el tiempo de ejecución. Mientas que para la GPU, aunque crecer el tamaño del problema implica un mayor consumo energético total, este sigue siendo muy inferior comparado con las soluciones paralelas en CPU.

5. Conclusiones y Trabajos futuros

El paralelismo busca acelerar el cálculo de los datos de las aplicaciones, ya que en algunos casos, los tiempos de ejecución obtenidos con los algoritmos secuenciales suelen ser inaceptables para los tiempos de respuesta requeridos. En particular, en el presente trabajo, se plantea como caso de estudio un problema de alta demanda computacional como N-body. La resolución del problema en su versión secuencial, obtiene tiempos de procesamiento muy elevados [20], y se han implementado soluciones paralelas, en dos arquitecturas de distinto tipo como CPU y GPU. Se ha logrado con éxito demostrar que se puede alcanzar una aceleración significativa con el uso de la GPU, para el caso de estudio.

Por otro lado, la reducción del consumo energético en los últimos años se ha convertido en un factor limitante en el uso de equipamiento tecnológico para la resolución de problemas del mundo real. Ya sea por lograr una escalabilidad de los sistemas en un futuro, como reducir los gastos económicos por el consumo de energía eléctrica, cada vez es mayor la importancia que se le está dando a la eficiencia energética de los equipos y de los algoritmos en el uso de los mismos.

Los resultados presentados, muestran la disminución del consumo energético al utilizar GPU por su alta aceleración de cómputo, con respecto a las versiones

paralelas en CPU. Para el caso de las soluciones paralelas implementadas para la CPU, se han conseguido mejores resultados en la versión de memoria distribuida en la que la ejecución se realiza en una sola máquina, comparada con la solución en memoria compartida. Sin embargo, la reducción del consumo energético total por la aplicación no es tan significativa entre estas dos soluciones comparadas con la solución en GPU.

Como trabajos futuros se plantea el uso de cluster de GPU y el estudio del modelo híbrido MPI-CUDA para su utilización, analizando los parámetros de escalabilidad y consumo. Asimismo, se busca generalizar la investigación a otras clases de aplicaciones.

Referencias

1. Silva Juliana M. N., Drummond Lúcia, y Boeres Cristina, "On Modelling Multicore Clusters", 22nd International Symposium on Computer Architecture and High Performance Computing Workshops, (2010).
2. Tinetti Fernando G. y Wolfmann Gustavo, "Parallelization Analysis on Clusters of Multicore Nodes Using Shared and Distributed Memory Parallel Computing Models", World Congress on Computer Science and Information Engineering, (2009).
3. Kirk David B. y Hwu Wen-mei W., "Programming Massively Parallel Processors: A Hands-on Approach", Morgan Kaufmann, (2010).
4. Piccoli María Fabiana, "Computación de Alto Desempeño en GPU", XV Escuela Internacional de Informática del XVII Congreso Argentino de Ciencia de la Computación, Editorial de la Universidad Nacional de La Plata, (2011).
5. Nvidia Corporation, "GPU gems", Pearson Education, (2003).
6. Song Jun Park, "An Analysis of GPU Parallel Computing", 2009 DoD Hogh Performance Computing Modernization Program Users Group Conference, publicado en IEEE, (2010).
7. Francis Kevin y Richardson Peter, Green Maturity Model for Virtualization, The Architecture Journal, págs. 9-15. (2008).
8. Schneider Electric, "Go Green, Save Green. The Benefits of Eco-Friendly Computing", (2008).
9. S.S. Verma, "GREEN COMPUTING". Departamento de fisica, SLIET, (2007).
10. Nicolaisen Nancy, "Green Computing with Intel® AtomTM Processor-Based Devices",http://software.intel.com/en-us/articles/green-computing-with-intel-tomprocessor-based-devices/, (2010).
11. amd.com/public, "Meeting the challenges of the future with innovative solutions for public sector IT needs", (2011).
12. J. Balladini, F. Uribe, R. Suppi, D. Rexachs, E. Luque. "Factores influyentes en el consumo energético de los Sistemas de Cómputo de Altas Prestaciones basado en CPUs y GPUs". Facultad de Informática, Universidad Nacional del Comahue, Argentina, y Departamento de Arquitectura de Computadores y Sistemas Operativos, Universidad Autónoma de Barcelona, España. XVII Congreso Argentino de Ciencias de la Computación, (2011).
13. Francisco Chinchilla, Todd Gamblin, Morten Sommervoll, Jan F. Prins , "Parallel N-Body Simulation using GPUs ", Department of Computer Science , University of North Carolina at Chapel Hill , http://gamma.cs.unc.edu/GPGP , Technical Report TR04-032, (2004).

14. Bruzzone Sebastian, “LFN10, LFN10-OMP y el Método de Leapfrog en el Problema de N Cuerpos”, Instituto de Física, Departamento de Astronomia, Universidad de la República y Observatorio Astronómico los Molinos, Uruguay, (2011).
15. Andrews Gregory R., “Foundations of Multithreaded, Parallel, and Distributed Programming”, Addison-Wesley, (2000).
16. Jeroen Bédorf, “High Performance Direct Gravitational N -body Simulations on Graphics Processing Units”, Universiteit van Amsterdam, (2007).
17. Nvidia, “NVIDIA CUDA C Programming Guide”, (2011).
18. Nvidia, “CUDA C BEST PRACTICES GUIDE”, (2012).
19. Perez Cristian y Piccoli M. Fabiana, “Estimación de los parámetros de rendimiento de una GPU”, Mecánica Computacional Vol XXIX, págs. 3155-3167, (2010).
20. Erica Montes de Oca, Laura De Giusti, Armando De Giusti, Marcelo Naiouf, “Comparación del uso de GPU y Cluster de multicore en problemas de alta demanda computacional”, XVIII Congreso Argentino de Ciencias de la Computación, CACIC 2012, pág. 267-275, (2012).

N-Body Simulation Using GP-GPU: Evaluating Host/Device Memory Transference Overhead

Sergio M. Martin[1], Fernando G. Tinetti[2,3], Nicanor B. Casas[1],
Graciela E. De Luca[1], Daniel A. Giulianelli[1]

[1]Universidad Nacional de La Matanza
Florencio Varela 1903 - San Justo, Argentina
[2]III-LIDI, Facultad de Informática, UNLP
Calle 50 y 120, 1900, La Plata, Argentina
[3]Comisión de Inv. Científicas de la Prov. de Bs. As.
fernando@info.unlp.edu.ar, {smartin, ncasas, gdeluca, dgiulian}@ing.unlam.edu.ar

Abstract. N-Body simulation algorithms are amongst the most commonly used within the field of scientific computing. Especially in computational astrophysics, they are used to simulate gravitational scenarios for solar systems or galactic collisions. Parallel versions of such N-Body algorithms have been extensively designed and optimized for multicore and distributed computing schemes. However, N-Body algorithms are still a novelty in the field of GP-GPU computing. Although several N-body algorithms have been proved to harness the potential of a modern GPU processor, there are additional complexities that this architecture presents that could be analyzed for possible optimizations. In this article, we introduce the problem of host to device (GPU) – and vice versa – data transferring overhead and analyze a way to estimate its impact in the performance of simulations.

Keywords: N-Body Simulation, GPU Optimization, Data Transference Overhead.

1 Introduction

The N-body problem is largely known within the physics, engineering, and mathematical research faculty. It is commonly used to calculate – as precisely as possible – the future position, velocity, momentum, charge, potential, or any other aspect of a massive/charged body in regard to other bodies that interact with it within a time interval. Although some efforts have been made [1], many theorists have unsuccessfully tried for centuries to find a purely mathematical solution that could resolve any application of this problem in a series of steps linearly related to the amount (n) of bodies. Therefore, currently, the only way to approximate to a real solution is to use a differential method with tiny time slices (differentials) using the power of modern computers. However, this approach presents some downsides.

First, the usage of finite (as opposed to infinitesimal) time differentials is detrimental to the precision of the result. All positions and momentums are taken

from the starting moment of the differential and are kept as constants during the calculation. Since the simulated forces remain constant during such differential, the results obtained suffer from a subtle degradation after each iteration. In consequence, the larger time differential is used, the more error is produced [2].

On the other hand, if we use smaller time differentials for the simulation, more iterations will have to be calculated until the end time is reached. As a result, simulations will require more computing time.

It is therefore important to keep in mind that the length of the selected time differential ultimately defines the precision admissible for the result expected, and the time that a computer will take to complete the simulation. Using high-precision libraries to augment the precision will also redound in increased computing time [3].

A way to calculate the amount iterations (n) to be simulated is to evaluate the inverse relation between the entire simulation time interval (Δ_t), and the time differential (∂t) as shown in Eq. (1).

$$n = \frac{\Delta_t}{\partial t} \tag{1}$$

Yet another reason why time differential is an important factor to be taken into account is that it defines the amount of data being transferred between the host memory – traditionally known as RAM – and the device – graphics processing unit – through the PCI-Express bus. The time taken for the simulation will increase if more resources/time should be spent on unnecessary data transmission rather than just processing [4].

In traditional CPU-based schemes, this kind of data transference overhead is negligible since all data is present and up-to-date within the host memory after each iteration is calculated. In those cases, it is possible to use/access to all the positions of all n bodies and use them in real-time – for instance, for saving them into an output file, or rendering them into the screen. However, when GPU devices are used for these algorithms it is required to define explicit data transferences from the results obtained within the device memory back to the host in order to enable them for any use. Such overhead is detrimental to the overall performance and any efforts made to reduce it can yield significant optimizations [5]. Estimating such overhead is the object of our analysis in this article.

2 CUDA Implementation of N-BODY

The CUDA programming model as an extension of the C language provides an excellent scheme to parallelize scalable N-Body algorithms for GP-GPU architectures [6]. In this model, in opposition to the conventional multicore CPU model, the programmer is encouraged to create as many threads as needed depending on the amount of data elements in the problem. By doing this, it is possible, in the generality of cases, to yield the maximum performance from the many simple yet extremely

parallelizable GPU cores. Of course, existing algorithms similar to the one used in this article are not exceptions [7] [8].

In the particular implementation of our N-body algorithm for CUDA, we created one thread per body in the simulation that will be in charge to execute the same function – called *CUDA kernel* – within the GPU processor. This kernel is programmed to execute the following steps:

- Load a single body's initial values from the device global memory. Each thread will load a different body based on its thread ID.
- For each other body in the simulation:
  - Load the body's values from the device global or shared memory.
  - Calculate the force that all other bodies impose to the loaded body.
- Save the new values for acceleration in the body data back into the device global memory.

Although threads perform better when no synchronization or communication functions are executed within the kernel, the CUDA architecture allows the programmer to specify *blocks* where a certain number of threads – depending on the infrastructure capability – can work coordinately. Based on this possibility, several memory access optimizations can be done in order to reduce the memory latency overhead.

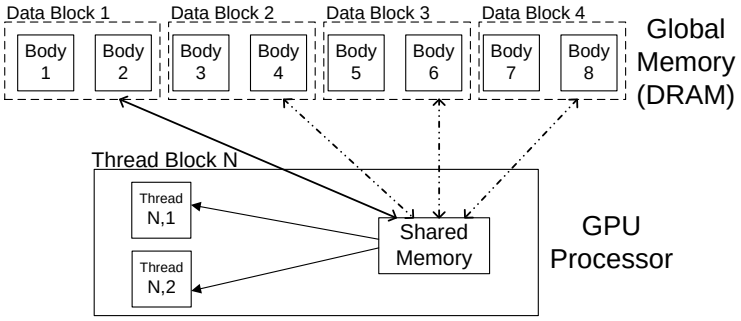


Fig. 1: Shared Memory utilization for the N-Body CUDA kernel

The most successful optimization that we implemented was the usage of intra-block shared memory. Since constantly accessing global memory (low latency) forces executing threads to stall until data is effectively loaded, the overall performance is greatly reduced. For that reason, this architecture provides the programmer with intermediate memory banks – such as shared and register memory – which reside within the processor and could be used to reduce the amount of accesses to global memory.

Fig. 1 shows an example of such optimization where threads are grouped into one-dimensional blocks of size two[1]. In the same fashion, bodies' data present in global memory were divided in data blocks that are loaded, one by one, into the block shared memory. By doing this, all threads read data from the global memory only once and

[1] This size was arbitrarily defined for simplicity reasons in this article while, in fact, blocks of 512 threads were actually used in our experiments.

access it several times within the shared memory, thus reducing the total memory latency overhead.

The pseudo-code shown next represents the N-body kernel to be executed by every thread using the CUDA terminology:

```
void nbodyKernel(vec)
{
  thread_body = vec[TID + BID * BSIZE]

  For each i in  BCOUNT Do
     Shared_data[TID] = vec[TID + BSIZE * i]
     For each j in BSIZE Do
       Compute(thread_body, i*BSIZE + j)
       Update acceleration of thread_body
     End for
  End For

  vec[TID + BID*BSIZE] = thread_body
 }
```

Where:

- **thread_body** is the private memory for the body data pertaining to each thread.
- **vec** is the collection of bodies' data stored in the global memory of the device.
- **Shared_data** is a vector of size **BSIZE** where a complete block of data is stored and used as shared memory by a particular block of threads.
- **TID** is the thread identifier within the block.
- **BID** is the block identifier.
- **BSIZE** is the size of each block.
- **BCOUNT** is the amount of blocks created.

A variety of other optimizations has been applied to the algorithm used in our experiments. Some of them have been already described in our previous work regarding the usage of multi-core clusters described in [9] and [10], and were used as the base for the CUDA version of our algorithm. However, more GPU-specific optimizations such as memory coalescing, shared memory usage, loop unrolling, interleaved thread-body processing were applied. Most of these optimizations are defined as good practices for any CUDA algorithm [11] [12]. Consequently, we assume for our experiments that the algorithm cannot be optimized any further.

3 Memory Transference Overhead

There are many types of research that requires scientists to run N-Body simulations in physics or engineering topics. In some, only the final result – for example, final position of the bodies involved – is needed; in others, it is more important to know the path that those bodies took during the simulation. Depending on each case – or a

combination thereof –, scientists could choose to have the intermediate results stored in a device, transmitted through a network or displayed on a screen. In other cases, they would discard part or the entire journey in order to reduce memory transference overhead.

As mentioned for CPU based algorithms, all information is present in the host memory to be used at all times. Even if it is not used, stored, or sent through a network during the simulation, no extra time is required for memory transmission. However, in the case of GP-GPU algorithms, copying the data back to the host is necessary if some action is to be performed with them.

It is important to mention that, even if no intermediate data is needed for the simulation purposes, it is still necessary to guarantee results with acceptable precision by calculating the necessary amount of iterations of rather small time differentials until the total simulation time is reached. This forces every simulation to be performed with a certain number of iterations, even if only the final result is needed.

In this research, we sought to measure the impact of data transmission on the overall performance of the algorithms, letting aside other possible overheads introduced by its usage. By measuring this, we were be able to determine how much performance can be gained by only obtaining the final results of a N-Body simulation, in comparison with transmitting the intermediate results at each iteration. This allowed us to define the minimum and maximum performance gain possible regarding data transmission between the host (CPU) and device (GPU), having all other possible combinations (for instance, transmitting one result every two iterations) in between those two results.

We have verified through experimentation that these relations do not vary when the iteration count[2] is changed. Using a rather high amount of iterations, deviation becomes insignificant. For iterations counts close to 1, however, execution interference from the operating system introduces a more noticeable deviation.

In order to measure how much overhead is introduced by transmitting data at each iteration in relation to doing so only at the beginning and the end of the simulation, we ran the same set of tests to compare two algorithms. Algorithm *Nbody1* transmits – yet it does not use – intermediate results after each iteration, and *Nbody2* calculates all iterations without interruptions. The architecture used for our tests is shown in Table 1, and the results obtained are shown in Table 2.

Table 1: GPU Architecture used.

GPU Device	GeForce GTX 550Ti
CUDA Cores	192
Capability	CUDA 2.1
DRAM	1 GB GDDR5

Table 2: GFlop/s obtained for both versions

n	*Nbody1*	*Nbody2*
4096	271	307
8192	315	333
16384	343	353
32768	357	362

The first detail to notice from the results is that the difference between the GFlop/s obtained from both versions – the amount of overhead introduced by data transmission – reduces as *n* (amount of bodies being simulated) increases. This can be

[2]We used 65536 simulation iterations in all our experiments.

explained by the fact that the algorithm complexity is quadratic – becomes 4 times bigger, when we double the data – while data transmission increases only linearly regarding the problem size – transmission time will only double. In other words, as the threads take more time to execute the kernel, the overhead of data transmission becomes less significant. This relationship can be seen in the results presented in Fig. 2.

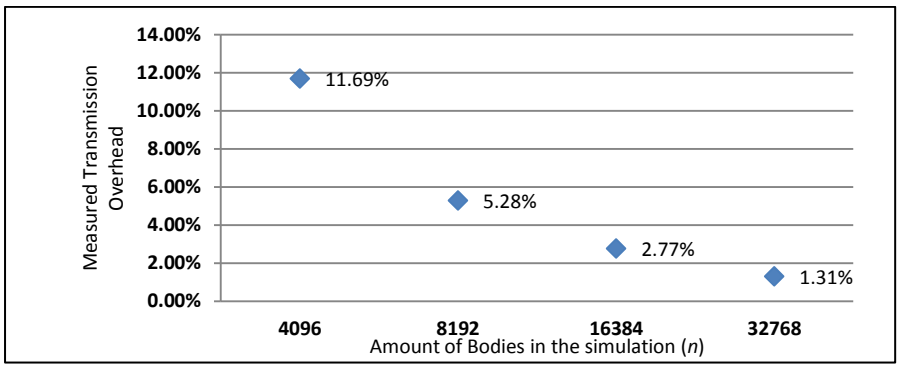


Fig.2: Measured transmission overhead ratio.

4 Transmission/execution ratio evaluation

Since we have empirically obtained values of ratio between the transmission overhead size (expressed in Flop) for several cases of *n*, we deemed necessary to look for a relationship that could allow us to evaluate this ratio for any given *n*. Moreover, expressing this relationship in terms of bytes and Flops could allow calculating an estimate of transmission overhead for other types of algorithms, and not only for N-body problems. The first step in order to obtain such relationship is to find how data transmission requirement increases given a discrete increase in one body. We used profiling tools [13] and techniques [14] that obtain precise information about memory usage directly from the hardware counters. Fig. 3 shows the host/device transference volume for a single iteration with *n*=4096 elements and Fig. 4 shows the host/device transference volume for the same N count.

	Source	Destination	Size (bytes)
1	Host Unpinned	Device	917504
2	Device	Host Unpinned	917504

Fig.3: Transmitted data for one simulation iteration.

	Function Name	Achieved FLOPS [1]: Single FLOP Count
1	nbodyKernel	20,400,472,064.00

Fig.4: Single-Precision Flop Count for the N-Body kernel per iteration.

We need now a way to tell how much data transmission increases when adding another body to the simulation. This can be obtained by multiplying the calculated data transmission per iteration by count of the iterations being simulated and dividing it by the *n* used. The resulting expression will determine *T(n)* – total transmission requirements – as a function of *n*. Eq. (2) shows n for the performed test:

$$T(n) = \frac{917504\ bytes}{4096}\ n = \ 224n\ bytes \tag{2}$$

The second step is to determine the size of the problem – expressed in MFlop – increases, given a similar increase in one element. To obtain this value, the same profiling tool allowed us to know how many floating operations were performed during the execution of the N-body kernel. As a result we can consider *F(n)* – total amount of Flop – as a function of *n*, using the obtained single-precision Flop count per body/body compute as in Eq. (3):

$$F(n) = \ 1216\ n(n-1)\ FLOP \tag{3}$$

Having *T(n)* and *F(n)* as functions of *n*, it is possible to establish the relationship between the bytes of data being transmitted and the amount of Flops for each additional element of an algorithm with quadratic complexity. As a result, we can obtain a data overhead ratio (*dor*) as in Eq. (4):

$$dor(n) = \frac{N(n)}{F(n)} = \frac{224\ n\ bytes}{1216\ \ n(n-1)\ FLOP} \cong \frac{0{,}185}{(n-1)} \left[\frac{byte}{FLOP}\right] \tag{4}$$

The data overhead ratio (*dor*) obtained indicates, for this algorithm, how many bytes will be transmitted per floating point operation to be executed, given *n* elements. The *dor* value for every integer between 4096 and 32768 resemble the same inverse relation that our experimental measures shown in Fig. 2.

What is most important about this relation is that it is architecture-independent. This means that, no matter which GPU device model we use, the execution of this kernel will have the same ratio between data transmission and Flop processing. Thus, we only have to link it with the actual cost of transmission of this specific architecture to get its fraction of the performance overhead.

This proportion can be easily calculated since we know that the optimal performance of the GPU device doesn't vary, and it is only being reduced by the data transmission overhead. Thus, we can assume that the increase in the problem size – measured in GFlop – is the *r* relation for the performance drop observed in Table 2. For N = 4096, Eq. (5) reflects this increase:

$$r(4096) = \frac{307 - 271}{307} = 0.117\ \frac{Bytes\ (transfered)}{FLOP\ (processed)} \tag{5}$$

Therefore, if this relation is observed for *n* = 4096, there has to be a constant *k* that allows to represent perfectly the percentage of performance drop due to data transmission as seen from our measurements for this specific architecture. Calculating it from the *r*(4096) ratio value, we obtained the result shown in Eq. (6):

$$r(n) = k * \ dor(n) = \frac{480}{(n-1)} \tag{6}$$

Having the relation *r* as a function of *n* will allow us to obtain the data overhead ratio for any *n* positive integer without having to perform any additional tests. As can be seen in Fig. 5, this inference matches perfectly with those measured in experiments and shown in Fig. 2.

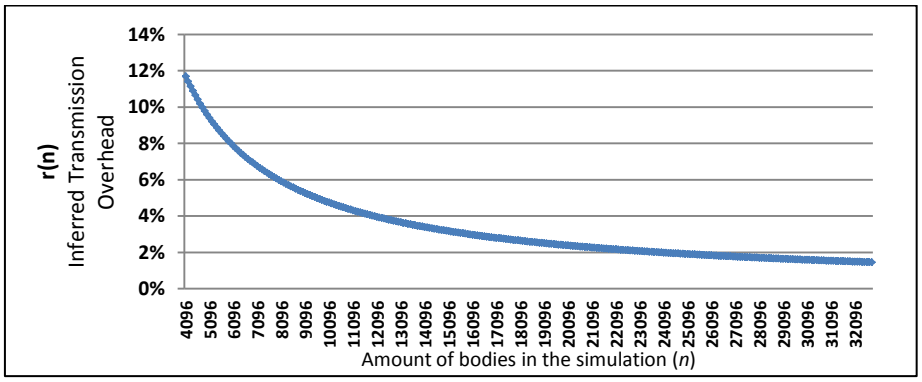


Fig.5: Inferred transmission overhead ratio.

It is important to note that the proportion *k* obtained is the particular value of the GPU device – and underlying architecture – we used. Therefore, for each other architecture used to execute our kernel, a new value for *k* should be provided that reflects the estimated performance drop.

On the other hand, since the *dor(n)* ratio will not vary between different architectures, it should be calculated only once per algorithm. Then, just combining it with the appropriate *k* proportion to obtain the *r(n)* of that specific algorithm-architecture scenario.

The most valuable aspect of having such pre-calculated proportions is that a table containing different *k* values for the available architectures, and *dor(n)* values for the available algorithms, we could predict the performance drop for data transmission for combinations of algorithms and architectures that were not tested in actual experiments.

4.1 Interleaved transference per iteration ratios

We have surmised through our tests that the performance overhead of transmitting the simulation's intermediate results at each iteration for different values of *n* can be estimated. However, it could also be helpful to calculate the overhead if just a certain portion of intermediate results should be gathered. In such case, we would have data transmissions every *m* number of iterations.

As we could appreciate in the previous section, the *dor(n)* ratio for this algorithm was calculated for a 1/1 proportion of transmissions per iteration. However, if we wanted to change that proportion to 1/2, (which means: transmitting every two iterations) its value would proportionately drop to a half. Thus, we can extend our definition of *r* to take into account the amount of iterations per transmission as in Eq. (7):

$$r(n,m) = \frac{k * dor(n)}{m} = \frac{480}{(n-1)m} \quad (7)$$

In order to test the accuracy of the estimations made with the *r(n,m)* equation, we verified its estimations with a series of tests using variations for the values of *n* and *m*. We confirmed that every result approached the estimations with negligible deviations. Therefore, such equation could effectively determine the impact of data transmission in a wide variety of cases. In Fig. 6, we show different curves as functions of *n*, using different values for *m*.

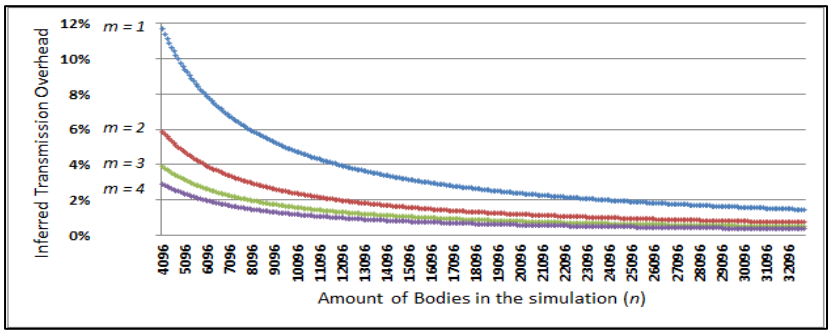


Fig.6: Inferred transmission overhead ratio for different values of *m*.

The extent at which scientists will be willing to sacrifice intermediate data to be discarded by this approach should be considered for each case. However, having estimations for all combinations of *n* and *m* we can provide valuable clues for establishing the best option in each case.

5 Conclusions and Further Work

Balancing and fine-tuning the two factors that define the numerical precision of a simulation (total time interval and differential) can be a very complicated task. Since they define the amount of iterations being calculated, they will also define how much real time will be spent on the actual calculations. Certainly, for scientists only interested in a final result, estimating the negligible data transference overhead is of a little interest. However, for simulations that need to store intermediate data, time spent on device/host transference would become an important issue.

Providing scientists with a way to estimate how much processing time will be added in data overhead – given the amount of iterations and the interleave transfers – could allow them to estimate the best option for their time/architecture availability without having to try all the possible combinations, which could demand more effort than the performing the simulation itself.

In that sense, we have defined and tested a method to estimate the impact of data transmission vs. processing time in GPU-based simulations and N-Body algorithms. It could be evaluated for other types of GP-GPU algorithms since we were able to narrow it down to a bytes/Flop relationship. We estimate that it would only require to

calculate a data overhead relation – a constant for the algorithm –, and a data transmission cost – a constant for the device, as a metric for size in Flops. However, more testing on a diversity of algorithms and architectures should be performed in order to verify whether this relationship could be extrapolated.

The next step on this research will be focused in evaluating how other device performance counters could best allow us to estimate the costs of transmitting data, and how it could be optimized. Additionally, it will be necessary to determine how to estimate the transference overhead N-body algorithms ran in multiple device architectures or GPU clusters. Those cases hold much larger penalties for data transferences, and thus offer more challenges for data overhead estimation.

References

1. F. Diacu, "The solution of the n-body problem",*The Mathematical Intelligencer*, 18(3), 1996, 66-70.
2. P. E. Zadunaisky, "A method for the estimation of errors propagated in the numerical solution of a system of ordinary differential equations" *Proceedings from Symposium on The Theory of Orbits in the Solar System and in Stellar Systems*. August, 1964. Thessaloniki, Greece.
3. T. Nakayama, D. Takahashi,"Implementation of Multiple-Precision Floating-Point Arithmetic Library for GPU Computing". *Parallel and Distributed Computing and Systems*. December, 2011. Dallas, United States.
4. C. Gregg, K. Hazelwood, "Where is the data? Why you cannot debate CPU vs. GPU performance without the answer".*IEEE International Symposium on Performance Analysis of Systems and Software.* April, 2011. Texas, United States.
5. D.Mudigere,"Data access optimized applications on the GPU using NVIDIA CUDA".Master's thesis, *Technische Universität München.*October, 2009. Munich, Germany.
6. J.Nickolls, I. Buck, M. Garland, K. Skadron, "Scalable parallel programming with CUDA". *Queue - GPU Computing*, 6(2), 2008, 40-53.
7. J. Siegel, J. Ributzka, Li Xiaoming, "CUDA memory optimizations for large data-structures in the Gravit simulator". *International Conference on Parallel Processing Workshops*. September, 2009. Vienna, Austria.
8. R. G. Belleman, J. Bedorf, S. P. Zwart, "High Performance Direct Gravitational N-body Simulations on Graphics Processing Units". *New Astronomy*, 13(2), 2008, 103-112.
9. F. G. Tinetti, S. M. Martin "Sequential optimization and shared and distributed memory parallelization in clusters: N-Body/Particle Simulation." *Proceedings of Parallel and Distributed Computing and Systems*.November, 2012. Las Vegas, United States.
10. F. G. Tinetti, S. M. Martin, F. E. Frati, M. Méndez. "Optimization and parallelization experiences using hardware performance counters". *International Supercomputing Conference Mexico.* March, 2013. Colima, Mexico.
11. NVIDIA CUDA™ Programming Guide Version 5.0, NVIDIA Corporation, 2012.
12. NVIDIA CUDA™ Best Practices Guide Version 5.0, NVIDIA Corporation. 2012.
13. NVIDIA Nsight™ Visual Studio Edition 3.0 User Guide. NVIDIA Corporation. 2013.
14. G.L.M. Teodoro, R.S. Oliveira, D.O.G. Neto, R.A.C. Ferreira, "Profiling General Purpose GPU Applications". *21st International Symposium on Computer Architecture and High Performance Computing.* October, 2009. Sao Paulo, Brazil.

Procesamiento de Señales SAR: Algoritmo RDA para GPGPU

Mónica Denham[1,2,3], Javier Areta[1,2], Isidoro Vaquila[2,5], Fernando G. Tinetti[3,4]

[1]Ingeniería Electrónica, Sede Andina, Universidad Nacional de Río Negro
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) Argentina
[3]Instituto de Investigación en Informática LIDI. Fac. de Informática – UNLP
[4]Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)
[5]INVAP
{mdenham,jareta}@unrn.edu.ar, ivaquila@invap.com.ar,
fernando@lidi.unlp.edu.ar

Resumen En este trabajo se presenta una solución secuencial y una paralela del algoritmo RDA (*Range Doppler Algorithm*) para el procesamiento de señales de radares SAR (*Synthetic Aperture Radar*). La solución paralela se desarrolló en C CUDA para GP-GPU (*General Purpose Graphic Processing Units*). Se describe la solución desarrollada, se muestran los primeros resultados y se describen las futuras optimizaciones para dicho algoritmo.

Keywords: Procesamiento Paralelo, HPC, CUDA, GP-GPU, Procesamiento de señales, radares SAR.

1. Introducción

Los radares de apertura sintética (SAR) son radares de pequeñas dimensiones que se acoplan a aeronaves (aviones o satélites) y aprovechan el desplazamiento de dicha aeronave para obtener imágenes de alta resolución. Se utilizan para formar imágenes de la superficie terrestre, detectar objetivos, realizar el seguimiento de objetivos móviles, etc.

En este trabajo la información recolectada y almacenada por el radar es procesada para formar imágenes de la superficie terrestre.

El modo de operación de estos radares se basa en que el radar avanza con la trayectoria de la aeronave mientras envía sucesivos pulsos y almacena sus ecos. Toda la información almacenada se sintentiza para formar una única imagen [3] [4] [9] [11] [12].

Debido a la frecuencia de envío de pulsos y al muestreo de los ecos recibidos, los datos recolectados son almacenados en arreglos de 2 dimensiones: las filas corresponden a cada pulso enviado (dimensión llamada acimut) y cada eco es muestreado y almacenado en una fila (el muestreo en rango define las columnas, dimensión llamada rango). Debido a las altas frecuencias con que se opera (en

rango y acimut), las matrices suelen ser de grandes dimensiones, normalmente almacenan millones de datos. A estos datos se los conoce como *datos crudos*.

En la actualidad, existen diversos algoritmos para el procesamiento de señales SAR [4] [6] [9] [11] [12]. Los algoritmos más conocidos y utilizados son: *Range Doppler Algorithm* (RDA), *Chirp Scaling Algorithm* (CSA), *Omega-K Algorithm*, *Back-projection Algorithm*, etc.

Dichos algoritmos se basan en aplicar filtros sobre los datos, transformadas de Fourier, antitransformadas, etc., todas operaciones con altos costos computacionales. Como se especificó anteriormente, estos algoritmos operan sobre una gran cantidad de datos. Estas características obligan a desarrollar soluciones desde la tecnología HPC (*High Performance Computing*). Además, es frecuente que este procesamiento esté ligado a aplicaciones de tiempo real haciendo aún más necesaria la implementación de algoritmos con bajos tiempos de respuesta, sin perder calidad de las imágenes generadas.

En este trabajo se utilizan procesadores gráficos de tipo GPGPU como arquitectura de ejecución paralela. Estas son placas que nacen para procesamiento gráfico (en el mercado de video juegos) pero su alto poder de cómputo, alto rendimiento y bajo costo ha hecho que se desarrollen placas gráficas de uso general (GPGPU: *General Purpose Graphic Processing Unit*).

Este trabajo es la continuación de [7], donde se enumeran las principales características del procesamiento de datos crudos SAR, como así también los pasos de los algoritmos RDA y CSA. Además, se muestran los rasgos más importantes de las arquitecturas GPGPU, poniendo énfasis en la organización de los componentes de las placas gráficas, la jerarquía de *threads* y jerarquía de memorias. Además se introduce las principales características de la programación en C CUDA.

En las siguientes secciones se presentan los aspectos principales de la solución secuencial y paralela del algoritmo RDA. La solución paralela es una primera aproximación en la que aún no se consideran aspectos más avanzados del hardware [2] [5]. Luego se presentan resultados obtenidos con dichos algoritmos y una comparación y análisis de los mismos. Además se plantean los pasos futuros que corresponden con optimizaciones del algoritmo.

1.1. Procesamiento de Señales SAR: Range Doppler Algorithm

El algoritmo RDA fue desarrollado en 1978 para procesar datos del primer radar SAR (SEASAT SAR) y hasta la actualidad es uno de los algoritmos más utilizados. Su principal característica es que usa operaciones en el dominio de la frecuencia en ambas dimensiones (rango y acimut), operando de esta forma en 1 dimensión, logrando mayor simplicidad [4].

El algoritmo se basa en la aplicación de tres operaciones: compresión en rango, RCMC (*Range Cell Migration Correction*), y compresión en acimut. Con estas operaciones se enfocan los datos en rango, luego se corrige migración de celdas en rango y por último se comprimen los datos en acimut para enfocar los datos en esta segunda dimensión.

La compresión de rango se lleva a cabo realizando una convolución rápida en cada fila de la matriz: se realiza una FFT para llevar los datos al dominio de la frecuencia, se multiplica por un filtro y por último se realiza una IFFT para llevar los datos al dominio del tiempo nuevamente.

Asumiendo el caso de radares transportados en aviones, donde las distancias son cortas, se puede asumir que la superficie terrestre es plana. Considerando un objetivo puntual, es facil observar que a medida que avanza el radar en su recorrido, la distancia del radar a dicho objetivo cambia con el tiempo (acimut). Esta diferencia en la distancia en función del tiempo en acimut genera migración de celdas en rango en los datos crudos. Es necesario entonces corregir esta migración de celdas. Dicha corrección se lleva a cabo mediante un corrimiento de celdas en función de la migración producida.

Por último, se realiza la compresión en acimut: cada columna de la matriz se pasa al dominio de la frecuencia (aplicando una FFT), se aplica un filtro, y se realiza una antitransformada para llevar los datos al dominio del tiempo.

2. RDA en C y C CUDA

Como punto de partida se ha desarrollado dicho algoritmo en MATLAB. Esta primer implementación aportó claridad y experiencia en las distintas operaciones del algoritmo y sus posibles implementaciones.

Además, dicha implementación permite verificar la correctitud de las operaciones a cada paso, haciendo que el desarrollo en C y C CUDA se desarrollen con plena seguridad.

El algoritmo secuencial en el lenguaje C se desarrolló con el objetivo de tener una métrica con la cual poder comparar el rendimiento del algoritmo paralelo.

Particularmente, se ha trabajado con una librería concreta para las operaciones de FFT e IFFT desarrollada por Mark Borgerding. Esta librería implementa de forma rápida las FFT e IFFT y se puede utilizar con datos de simple o doble precisión (`http://sourceforge.net/projects/kissfft/`).

Además de los algoritmos secuenciales se ha desarrollado el algoritmo en paralelo, especialmente diseñado e implementado para su ejecución en GPGPU.

Paralelización de RDA

Actualmente se dispone de una primer versión de RDA en C CUDA, la cual será explicada y evaluada. Este primer desarrollo de la solución paralela no se considera que sea óptima, pero constituye un punto de partida para llegar a soluciones más eficientes.

En el corto plazo, se buscarán modificiones al algoritmo paralelo para, principalmente, realizar una distribución de trabajo eficiente y una utilización de la jerarquía de memorias que logre rendimiento óptimo.

Como ya se ha expuesto, muchas de las operaciones de RDA se basan en aplicación de filtros, convoluciones, FFT, IFFTs. La paralelización del algoritmo se basa en: las operaciones FFT e IFFT se resuelven utilizando las operaciones

paralelas de la librería `cuFFT` y las aplicaciones de filtros, convoluciones, normalizaciones, se paralelizan creando bloques de *threads* y haciendo que cada *thread* compute un único elemento del arreglo (señal).

Libreria *cuFFT* [8] es una librería desarrollada por NVIDIA [1] que resuelve las Transformadas Rápidas de Fourier (y antitransformadas). Las operaciones en esta librería se implementan con la estrategia *divide y vencerás* y logran algoritmos eficientes para conjuntos de valores reales y complejos. La librería `cuFFT` provee interfaces simples que permiten al usuario hacer uso de la potencia de cálculo y poder de paralelismo de las placas GPU, de forma transparente.

2.1. Compresión en Rango

Compresión en rango secuencial Esta operación se desarrolla de forma secuencial, fila por fila, realizando una convolución rápida: FFT, aplicación de filtro, IFFT. Como ya se ha expuesto, las operaciones FFT e IFFT se resuelven de forma secuencial con operaciones de la librería KISS_FFT. A su vez, el filtrado de la señal se resuelve con un producto punto también secuencial. A continuación se muestra un pseudocódigo simplificado de esta operación.

Compresión de Rango Secuencial

```
...
for (cada fila)
{
    FFT;
    producto punto;
    IFFT;
}
```

La transformada discreta de Fourier (TDF) tiene orden de complejidad $O(n^2)$. A su vez, las implementaciones FFT (*Fast Fourier Transform*) son algoritmos recursivos que siguen la estrategia *divide y vencerás*, y pueden llegar a reducir este orden de complejidad a $O(n \log n)$. El producto punto tiene orden de complejidad $O(n)$. Estas operaciones se ejecutan por cada una de las filas, llegando a un orden de complejidad que $O(n^2 \log n)$.

Teniendo en cuenta las dimensiones de las matrices de datos crudos, decrementar este orden de complejidad es un requisito fundamental para mejorar el rendimiento de esta aplicación.

Compresión en rango paralelo En esta primer propuesta paralela, el procesamiento también se realiza por cada una de las filas de la matriz: las operaciones FFT e IFFT se realizan utilizando operaciones de la librería *cuFFT* de CUDA. Como ya se ha dicho, las operaciones de dicha librería están optimizadas para obtener máximo rendimiento en GPU.

Luego de la FFT, se realiza un producto punto de toda la fila. Para dicha operación se implementa un *kernel* el cual se ejecuta con tantos *threads* como elementos tenga la fila. Todos los *threads* se ejecutan de forma concurrente (con un elevado paralelismo) y cada uno resuelve un único producto y lo almacena en un lugar de la fila resultante (figura 1).

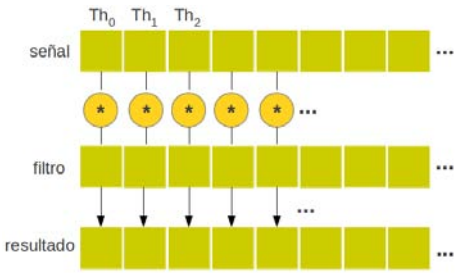


Figura 1. Kernel CUDA: cada *thread* calcula el producto de un elemento del vector resultante.

Una de las características de la utilización de GPUs y CUDA es que el programador puede crear un número de *threads* muy elevado, asumiendo que no hay restricción máxima de cantidad de *threads*. Por esto mismo, crear un *thread* por cada elemento de la fila, que compute un único producto es natural, independientemente de las grandes dimensiones que pueda tener la estructura de datos con la que se opera. De esta forma, todos estos productos se resuelven "de forma simultánea" y este es un de los aspectos principales de las placas gráficas.

Como se mencionó anteriormente, esta primer propuesta paralela realiza estas operaciones por cada una de las filas. A futuro, se preve implementar el procesamiento de todas las filas a la vez, aprovechando la ausencia de dependencia entre los datos y su procesamiento.

2.2. Range Cell Migration Correction (RCMC)

Por cada fila de la matriz se calcula la cantidad de celdas enteras a corregir debido a la migración. Además se calcula la migración a nivel de fracción de celda, como se propone en [4].

RCMC Secuencial Por cada fila, se calcula la cantidad de celdas que el objetivo haya migrado en los datos crudos, dependiendo del tiempo en acimut. Luego se realiza un corrimiento de todos los elementos de la fila en función del cálculo previo.

A su vez, se ajusta la migracion fraccionaria utilizando un kernel de interpolación propuesto en [4]. La fila se convoluciona con dicho kernel para realizar un ajuste más preciso. A continuación se muestra un pseudocódigo de RCMC.

Corrección de migración de celdas en rango secuencial

```
...
for (cada fila)
{
    cálculo de migración de celdas enteras;
    cálculo de migración de fracción de celdas;

     /* corrección de migración de celdas */
    corrimiento de toda la fila;
    convolución con el kernel de interpolación;
}
```

RCMC Paralelo A nivel de celda entera se realiza el corrimiento con un *kernel* paralelo: cada uno de los *threads* copia un dato en la fila teniendo en cuenta el corrimiento. Se copian en paralelo todos los datos de la fila, evitando una iteración sobre la misma.

La corrección de migración de celdas a nivel de fracción de celdas se ejecuta en paralelo teniendo un *thread* por elemento de la fila resultado y cada *thread* realiza la multiplicación y suma correspondiente. Nuevamente, se evita iterar sobre los elementos de la fila.

De forma similar a la operación anterior, se observa que una posible mejora es implementar todos los corrimientos de todas las filas de forma concurrente. Esto evita iterar sobre cada una de las filas.

2.3. Compresión en Acimut

Por cada una de las columnas de la matriz resultante de las dos operaciones anteriores se realiza: FFT de la columna, producto punto entre la columna y el filtro (el mismo se encuentra en dominio de la frecuencia) y por último IFFT del resultado, para llevar los datos resultantes al dominio del tiempo.

Compresión en Acimut secuencial Esta operación es similar a la compresión de rango, pero se trabaja columna a columna. Se trabaja en el dominio de la frecuencia, por lo que el filtro se transforma al comienzo del procesamiento, y cada columna se transforma antes del producto.

Compresión en acimut secuencial

```
...
for (cada columna)
{
    FFT columna
    producto punto columna y filtro
    IFFT resultado
}
```

Compresión en Acimut paralelo Nuevamente, esta operación aprovecha las operaciones de la librería cuFFT, las cuales garantizan operaciones eficientes para resolver las transformadas (antitransformadas) de Fourier. Por otro lado, el producto punto se resuelve en paralelo, utilizando un *thread* por cada elemento del vector.

Esto se realiza de forma secuencial sobre cada una de las columnas de la matriz, lo cual se prevee modificar para realizar de forma concurrente todas las columnas (no existe dependencia de datos).

En las próximas secciones se mostrarán y analizarán los primeros resultados obtenidos con ambas implementaciones.

3. Resultados

Se realizará la comparación de rendimiento del código secuencial y el rendimiento de la primer implementación paralela.

La arquitectura utilizada está compuesta por un host multicore de 4 procesadores Intel(R) Core i5-2500 CPU @ 3.30GHz, con 7.7GB de memoria y sistema operativo Ubuntu de 64 bits. En dicha CPU se encuentra conectada una placa gráfica de tipo GeForce GTX 550Ti con 192 cores CUDA (dispuestos en 4 multiproceadores) de fabricante NVIDIA.

Se ha experimentado con 3 imágenes: un objetivo puntual (imagen sintética: un único objetivo en el medio de la imagen), y 2 fotos reales, una de las cuales es la vista de un aeropuerto y otra de un crater de un volcán.

Para disponer de los datos crudos de dichas imágenes se ha utilizado un simulador que simula el proceso de generación de los datos crudos. Dicho simulador fue desarrollado por uno de los miembros del equipo de trabajo.

Como ejemplo, la figura 2 muestra la foto del aeropuerto utilizada y la imagen obtenida luego de procesar los datos crudos con el algoritmo RDA implementado.

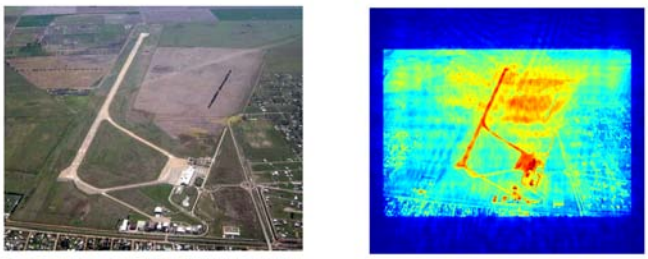


Figura 2. Imagen de aeropuerto, foto aérea e imagen obtenida con el algoritmo RDA.

Para los 3 casos de experimentación, se han tenido en cuenta los siguientes valores: el tiempo de exposición del objetivo es 3.4 segundos, y se trabaja a una frecuencia de envío de pulso de 600Hz. En total se toman 2000 pulsos.

Por otro lado, el eco del pulso recibido se muestrea para su almacenamiento y posterior procesamiento (valores discretos). La frecuencia de muestreo es de 120MHz. Debido a la distancia que se desea cubrir en rango, y al resto de valores propios de la geometría SAR, se almacenan 8000 datos en rango. Esto es, el eco recibido se muestrea y se obtienen 8000 valores. Estos valores definen matrices de 2000 filas de 8000 muestras (columnas). Esto corresponde a la matriz de datos crudos, matrices intermedias e imagen enfocada (imagen final).

Como el algoritmo RDA está bien definidos en 3 operaciones bien visibles, los primeros análisis se han realizado sobre cada una de estas operaciones. Se obtuvieron los tiempos de ejecución de las operaciones: compresión en rango, RCMC y compresión en acimut.

Los tiempos obtenidos se muestran en el cuadro 1 para las 3 operaciones del algoritmo RDA (en segundos). Se puede observar que los tiempos son muy similares para las 3 imágenes, por lo que se expondrán dichos tiempos en figuras sólo para un caso, valiendo el análisis para el resto. En dicho cuadro, RC es compresión en rango, RCMC es corrección de migración de celdas en rango y AC es compresión en acimut. Por cada imagen se muestra los tiempos secuenciales (Sec) y los tiempos del algoritmo paralelo (Par).

	Objetivo Puntual		Aeropuerto		Volcan	
	Sec	Par	Sec	Par	Sec	Par
RC	97.32	4.88	97.20	4.92	97.35	4.91
RCMC	0.57	0.04	0.56	0.04	0.57	0.04
AC	4.64	0.68	4.65	0.68	4.65	0.68

Cuadro 1. Tiempos obtenidos para las 3 imágenes (en segundos).

Para continuar con el análisis, la figura 3 muestra los tiempos del algoritmo RDA con una de las imágenes (objetivo puntual).

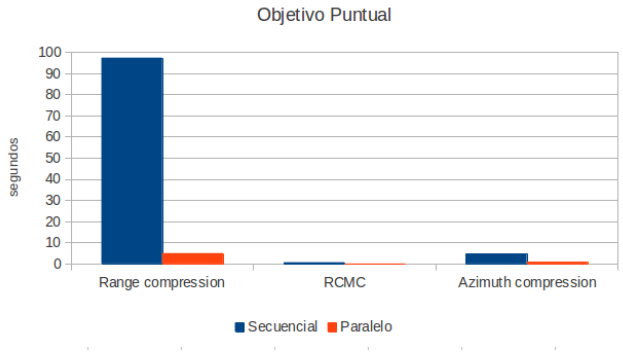


Figura 3. Tiempos de ejecución del algoritmo RDA con la imagen del objetivo puntual.

En dicha figura se puede distinguir con exactitud la ganancia de paralelizar la compresión en rango, pero las otras dos operaciones no se puede determinar debido a la escala del eje Y (tiempos en segundos). Por esto se muestran en la figura 4 estas operaciones en gráficos distintos, para poder comparar de forma precisa la diferencia de tiempos del algoritmo secuencial y paralelo. Considerar el cambio de escala del eje Y para dichos gráficos.

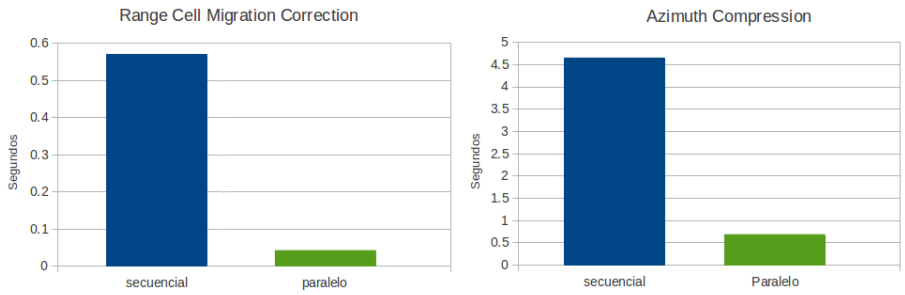


Figura 4. Tiempos de las operaciones RCMC y compresión en acimut en detalle.

En todos los casos es posible observar que la compresión en rango es la operación que más tiempo requiere cuando se resuelve de forma secuencial. Dicha operación resuelve una FFT, un producto punto y una IFFT por cada una de las filas. Debido a la operación FFT que se realiza, se realiza *zero padding* a los datos (filas), por esto se trabaja con 16000 valores en rango. Al finalizar esta operación, se vuelven a seleccionar los datos centrales de la antitransformada, que es donde se concentra la información útil (y se sigue el procesamiento usando 8000 muestras en rango).

Lo anteriormente dicho no sucede en las operaciones de corrección de celdas (se trabaja con filas de 8000 complejos) ni en compresión en acimut (columnas de 2000 complejos).

Tomando los tiempos secuenciales de cada operación como referencia, para la compresión de rango se observa un 95 % de reducción en el tiempo, en la operación RCMC un 93 % de reducción y en la compresión en acimut se observa un 86 % de reducción (aproximadamente).

Como se ha mencionado anteriormente, la solución paralela propuesta es una primer aproximación, para la cual no se consideran aspectos fundamentales en la optimización de códigos CUDA: cantidad y jerarquía de *threads* en cada kernel, movimiento de datos a memorias de más rápido acceso (local, compartida, textura o constante) [2] [5]. El próximo paso a seguir en esta línea es modificar estos aspectos y así lograr algoritmos paralelos más eficientes.

4. Conclusiones

Este trabajo presenta los primeros resultados del algoritmo RDA secuencial y paralelo utilizando C CUDA en GPGPU.

El algoritmo RDA cumple con los requerimientos necesarios para que su implementación y ejecución en GPU sea conveniente: alta carga computacional, independencia de datos, mínima transferencia de datos entre CPU y GPU (solo al comienzo y al final del procesamiento), no existen secciones críticas, los datos se mantienen en matrices logrando mapear las estructuras de datos con la disposición de los *threads* en los grids en cada kernel.

Los primeros resultados obtenidos muestran que la paralelización del algoritmo logra reducir significativamente los tiempos en las 3 operaciones del algoritmo RDA. Esto muestra que la paralelización es efectiva, y, teniendo en cuenta posibles optimizaciones, el algoritmo RDA promete un muy buen rendimiento en GPGPU. A corto plazo se espera optimizar estas tres operaciones para obtener algoritmos más eficientes aún.

Este trabajo muestra un primer paso abordando la comparación del algoritmo secuencial y del algoritmo paralelo. Como trabajo futuro también se estima la posibilidad de evaluar algoritmos propuestos en la literatura, con el fin de realizar comparaciones entre distintos algoritmos paralelos.

Referencias

1. NVIDIA home page, `http://la.nvidia.com/page/home.html`
2. Farber, R.: CUDA Application Design and Development. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edn. (2011)
3. Hein, A.: Processing of SAR Data. Springer (2004)
4. Ian G. Cumming, F.H.W.: Digital Processing of Synthetic Aperture Radar Data. Artech House (2005)
5. Kirk, D.B., Hwu, W.m.W.: Programming Massively Parallel Processors: A Hands-on Approach. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edn. (2010)
6. Meier, O.F..E.H., Nüesch, D.R.: Processing sar data of rugged terrain by time-domain back-projection. SPIE 5980, SAR Image Analysis, Modeling, and Techniques VII, 598007 (2005), `http://dx.doi.org/10.1117/12.627647`
7. Denham M., Areta J.: Procesamiento de señales sar en GPGPU. In: CACIC 2012 (2012), `http://sedici.unlp.edu.ar/handle/10915/23633`
8. NVIDIA: CUDA Toolkit cuFFT Library. Programming Guide (March 2011)
9. Richards, M.A.: Fundamentals of Radar Signal Processing. McGraw-Hill (2005)
10. Sanders, J., Kandrot, E.: CUDA by Example: An Introduction to General-Purpose GPU Programming. Addison-Wesley Professional, 1st edn. (2010)
11. Skolnik, M.I.: RADAR Systems. Tata McGraw-Hill
12. Soumekh, M.: Synthetic Aperture Radar Signal Processing with MATLAB Algorithms. Wiley-Interscience (1999)

Parallel implementation of a Cellular Automata in a hybrid CPU/GPU environment

Emmanuel N. Millán[1,2,3], Paula Cecilia Martinez[2], Verónica Gil Costa[4], Maria Fabiana Piccoli[4], Marcela Printista[4], Carlos Bederian[5], Carlos García Garino[2], and Eduardo M. Bringa[1,3]

[1] CONICET, Mendoza
[2] ITIC, Universidad Nacional de Cuyo
[3] Instituto de Ciencias Básicas, Universidad Nacional de Cuyo, Mendoza
[4] Universidad Nacional San Luis, San Luis
[5] Instituto de Física Enrique Gaviola, CONICET
{emmanueln@gmail.com,ebringa@yahoo.com}
http://sites.google.com/site/simafweb/

Abstract. Cellular Automata (CA) simulations can be used to model multiple systems, in fields like biology, physics and mathematics. In this work, a possible framework to execute a popular CA in hybrid CPU and GPUs (Graphics Processing Units) environments is presented. The inherently parallel nature of CA and the parallelism offered by GPUs makes their combination attractive. Benchmarks are conducted in several hardware scenarios. The use of MPI /OMP is explored for CPUs, together with the use of MPI in GPU clusters. Speed-ups up to 20 x are found when comparing GPU implementations to the serial CPU version of the code.

Keywords: General purpose GPU, Cellular Automata, multi-GPU

1 Introduction

Multicore CPUs have become widely available in recent years. As an alternative, Graphics Processing Units (GPUs) also support many cores which run in parallel, and their peak performance outperforms CPUs in the same range. Additionally, the computational power of these technologies can be increased by combining them into an inter-connected cluster of processors, making it possible to apply parallelism using multi-cores on different levels. The use of GPUs in scientific research has grown considerably since NVIDIA released CUDA [1]. Currently, 43 supercomputers from the Top 500 List (June 2013, `www.top500.org`) use GPUs from NVIDIA or AMD. Currently the most commonly used technologies are CUDA [1] and OpenCL [2]. Recently Intel entered the accelerator's market, with the Xeon Phi [3] x86 accelerator, which is present in 12 supercomputers from the Top 500 List (June 2013, `www.top500.org`).

This work evaluates the trade-off in the collaboration between CPUs and GPUs, for cellular automata simulations of the Game of Life [4]. A cellular

automaton (CA) [5] is a simple model represented in a grid, where the communication between the grid points or cells is limited to a pre-determined local neighbourhood. Each cell can have a number of finite states which change over time, depending on the state of its neighbours and its state at a given time [6]. Even though such a model is simple, it can be used to generate complex behaviour. CA have been used to implement diverse systems in different fields of science: biological systems [7], kinetics of molecular systems [8], and clustering of galaxies [9]. CA have also been implemented in GPU architectures in biology to simulate a simple heart model [10], infectious disease propagation[11], and simulations of laser dynamics[12].

The Game of Life has already been extensively studied [13][14][15]. In this paper, the model is used as a starting point to developed future CA simulations in hybrid (CPU+GPU) environments, as it has already been achieved for Lattice Boltzmann Gas simulations [16], the Cahn-Hilliard equation [17][18] and reaction-diffusion systems [19] [20]. A relatively small code was developed to run in multiple parallel environments, and the source code is available in the website of the authors (`https://sites.google.com/site/simafweb/proyectos/ca_gpu`). Five implementations of the Game of Life code with C were developed: one serial implementation which executes in a single CPU core, and four parallel implementations, including: shared memory with OpenMP, MPI for a CPU cluster, single GPU, and Multi-GPU plus MPI for a CPU-GPU cluster. The paper is organized as follows. Section 2 describes the Game of Life in general; Section 3 contains details of the code implementation; Section 4 includes code performance; and conclusions are presented in Section 5, including possible future code improvements.

2 Description of the Problem

The Game of Life, through a set of simple rules, can simulate complex behaviour. To perform the simulations of this work, a 2 dimensional (2D) grid with periodic boundaries is used. The grid cells can be "alive" or "dead", and a Moore neighbourhood (8 neighbours) [6] is considered for each cell. According to the cell state at time t and the state of its 8 neighbours, the new state for the cell at time $t+1$ is computed. The update is done simultaneously for all cells , which means that their change will not affect the state of neighbour cells at $t+1$ time. Time t+1 is often referred to as the "next generation". The following rules define the state of a cell in the Game of Life[4]:

- Any living cell, with less than two living neighbours will die in *t+1*.
- Any living cell, with two or three living neighbours will live in *t+1*.
- Any living cell, with more than three living neighbours will die in *t+1*.
- Any dead cell, with exactly three living neighbours will be alive in *t+1*.

3 Implementation

3.1 Serial Code Implementation

The serial implementation which executes the entire simulation in one CPU processor was used to verify the correct functioning of other implementations. The serial code is quite simple. It begins initializing the main 2D grid, size NxN, with zeros and ones randomly placed. N has to be an even integer. A secondary NxN 2D grid is used to store the states of each cell for the next iteration. Any random number generator can be easily used, but the standard *rand(seed)* function was used in all codes. For testing purposes, the seed used to generate the random numbers is always the same. The code runs up to a maximum number of iterations defined by the user. Within each iteration, every cell and their neighbourhoods are monitored, the four rules for the CA are applied, and the next set of states are stored into the secondary array. Once all cells are analysed, the secondary grid is copied into the main 2D grid, concluding a given iteration. Every m iterations, with m a pre-set integer number, the state of the main 2D grid is written to an output file. Since this is a simulation with periodic boundaries, care must be taken when the values of the neighbours of a cell have to be monitored. When the simulation ends, the program prints on the screen the amount of RAM memory used and the total execution time, separating by: filling time of the 2D grid, evolution time, and write-up time of output files.

3.2 OpenMP Implementation

The parallel shared memory implementation of the code was developed with OpenMP [21]. It uses two compiler directives (pragma) to execute in parallel the *for* loops in the serial implementation of the code. These two loops are the loop that iterates over the 2D grid applying the game's rules, and the loop which copies the secondary grid to the main grid. Each *pragma* was configured to use dynamic or static scheduling, with the work assigned to each thread defined by the type of scheduling method. When using static scheduling, each thread is assigned a number of *for* loop iterations before the execution of the loop begins. When using dynamic scheduling, each thread is assigned some portion of the total (chunk) of iterations, and when a thread finishes its allotted assignment returns to the scheduler for more iterations to process.

3.3 MPI Implementation

Using a cluster of computers to run the case of message passing with MPI significantly increases the complexity of the code, especially when dealing with periodic boundary conditions. To handle part of the communications, the strategy by Newman [22] was used. The initial grid is loaded as in the serial case, then this grid is divided into equal blocks, with block size given by the full grid size and the number of processors in a square topology, e.g. running in 4 processors will give a 2x2 topology. Then, each block is sent to a different processor.

When an iteration begins, each process exchanges its grid boundaries with its neighbouring processes. From the code by Newman [22], the functions *exchange(), pack()* and *unpack()* were used as basis for sending the boundaries and corners of each block from each process to its neighbours. After performing the boundary exchange, the number of living neighbours for each cell in each process is calculated, the block corresponding to that process is updated, and a new iteration begins, sending updated boundaries between neighbouring processes. When it is necessary to copy the state of the grid to a file, the block from each process is brought to the master process to be part of a single grid, which is then written. Because of the simple chosen communication scheme, the current code cannot deal with odd number of processes.

3.4 GPU Implementation

The implementation of the code for a single GPU uses NVIDIA CUDA and takes the CPU serial code as basis. Two kernels (C code functions that run on the GPU) were developed: one kernel controls the state of the neighbours and modifies the secondary grid for the next iteration; the other kernel is responsible for updating the main grid with the data obtained by the first kernel (it copies the secondary grid into the main grid). There has to be a blocking step to ensure that all the threads have finished executing their instructions within the same kernel for a given iteration, making two separate kernels necessary. In order to ensure that instructions were properly executed, it was necessary to place a barrier outside the first kernel, in the CPU host code. Is worth mentioning that there is extra communication time: to copy the input grid generated in the CPU to the GPU, and to copy the system grid to the main memory each time the grid is written to a file, because it is necessary to copy from the GPU global memory to the CPU main memory.

3.5 Multi-GPU Implementation

The parallel code with MPI and CUDA is similar to the parallel MPI implementation explained in section 3.3 and the GPU implementation from section 3.4. MPI sends a block of the principal grid to each node for processing, then each of these nodes run the simulation on a GPU. Two kernels were added: the first kernel prepares data to be sent from a GPU to neighbouring CPU nodes, and the second kernel is responsible for loading data received from neighbouring processors into the GPU. These kernels replace the functions *pack()* and *unpack()* of the CPU MPI code. This multi-GPU code also outputs the time to copy data between the CPU and GPU at each time step, and it was called "Halo time" as in [23]. This only takes into account data transfer between CPU RAM memory and GPU global memory, regardless of MPI communication time between nodes

4 Benchmarks

4.1 Infrastructure

Simulations were executed in three different environments.

- Workstation Phenom: 2.8 GHz AMD Phenom II 1055t 6 cores with 12 GB DDR3 of RAM memory. NVIDIA Tesla c2050 GPU, with 448 CUDA cores working at 1.15 GHz, and 3 GB memory. Slackware Linux 13.37 64 bit operating system with kernel 2.6.38.7, OpenMPI 1.4.2, Cuda 5.0 and gcc 4.5.3.
- Workstation FX-4100: 3.4 GHz AMD FX-4100 x4 with 8 GB of DDR3 RAM memory. NVIDIA GeForce 630 with 48 CUDA cores working at 810 MHz,and 1 GB of memory. Slackware Linux 14 64 bit operating system with kernel 3.2.29, OpenMPI 1.7 beta, Cuda 4.2 and gcc 4.5.2.
- Cluster Mendieta at the Universidad Nacional de Córdoba: 8 nodes with two 2.7 GHz Intel Xeon E5-2680 CPU with 32 GB of memory, 12 NVIDIA Tesla M2090 GPUs with 6 GB GDDR5 (177 GBps) of memory and 512 1.3 GHz CUDA cores. The connection between nodes is at 20 Gbps InfiniBand DDR, with the switch using star topology. With Linux CentOS 6.4, MPICH 3.0.4 and Cuda 5.0.

Since the Phenom and FX-4100 workstations have only a single GPU, the code developed with MPI + GPU executes various MPI processes in the same workstation and executes the same number of independent processes in a single GPU. Because of this, the communication time between MPI nodes through the network is not evaluated for these two environments.

4.2 Simulation Results: Analysis and Discussion

Simulations were performed for different grid sizes, for the same number of iterations (1000) in all cases. The output of the last iteration performed for all parallel codes was checked against the serial version to verify the correct operation of the parallel codes. All codes were compiled with optimization -O3. Grids were NxN, with $N = 500, 1000, 2000, 4000$, and 6000. Four processors were used for the parallel implementations using OpenMP, MPI, and Multi-GPUs. Simulation times for different runs with the same configuration vary only by few percent. In figure 1 it can be seen the evolution of a small region of a particular simulation. In these frames it can be seen a complex pattern drawn from the simple set of rules that are part of the Game of Life.

The performance tests performed on the Phenom workstation can be seen in tables 1 and 2. For the smallest dimension (N=500), the code in OpenMP, MPI and GPU is always faster than the serial version. For the Multi-GPU code the cost of copying the "halo" between different GPU processes is high and performance degrades. For all other simulations, parallel codes are always faster than the serial code. The OpenMP code was executed in four independent threads and, therefore, it was expected that the simulation time would decrease approximately four times. This was not the case, because the parallelization with

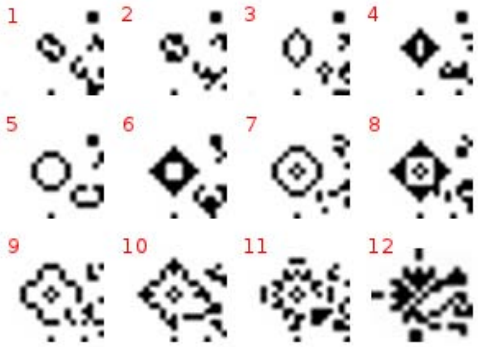


Fig. 1. Game of life simulation, for a grid with N=500, showing the evolution of a small region of the grid, for 12 different frames.

OpenMP was done by grid rows, and the use of the cache is far from optimal. The average speedups vs. the serial version, for all sizes, was 2.7 x. The implementation using MPI in 4 processors is approximately six times faster than the serial implementation, for all dimensions. The MPI code makes a division into blocks of the grid, and it is executed for the cases shown in tables 1 and 2, on a single workstation. Implementing OpenMP with blocks would likely increase performance. Speedups are given in table 5. When a single process is run on the GPU, the average speedup obtained for all grid sizes is 13 x. In the case of parallel Multi-GPU, the average speedup is 9.1 x. The copy time between GPU processes considering the smaller cases of the grid (500, 1000 and 2000) is greater than the computational time in each process. Therefore, running on a single GPU turns to be faster than using Multi-GPU for the grid sizes considered in this work. On the Phenom workstation (which has only one GPU), the Multi-GPU code executes multiple independent processes in the same GPU, at the same time. The largest case (N=6000) executes faster in multiple processes on the same GPU (Multi-GPU) than in a single process in the GPU, giving a 17 x speedup over the serial version. This improvement is due to the way in which the GPU scheduler administers the execution of the threads [24]. In addition, domain decomposition provides data locality, improving memory latencies [24] [25].

Tests performed on the cluster "Mendieta" were the same tests performed in the Phenom workstation. In the case of the cluster "Mendieta", two nodes with two GPUs in each node were used. All tests were performed using one GPU per node, except for the largest case with N = 6000, which was also executed for two GPUs per node.

"Mendieta" was being shared with others users at the time of testing. This might be the reason why some execution times were greater than when using the Phenom workstation, which was used exclusively for these tests. Tables 3 and 4 contain the results. There are large differences in communication time between the multi-GPU case in Mendieta and the Phenom workstation for the transfer

Table 1. Simulation time in seconds for CPU serial, OpenMP and MPI codes, executing in the Phenom Workstation.

N	CPU Serial				OpenMP				MPI			
	fill	evolve	output	total	fill	evolve	output	total	fill	evolve	output	total
500	0.003	2.25	0.05	2.3	0.003	0.93	0.05	0.99	0.004	0.22	0.05	0.28
1000	0.01	10.03	0.17	10.21	0.01	3.44	0.19	3.64	0.02	1.15	0.21	1.37
2000	0.04	39.85	0.66	40.55	0.04	14.55	0.74	15.33	0.07	7.84	0.88	8.79
4000	0.16	167.15	2.64	169.95	0.16	56.75	2.95	59.86	0.25	30.23	3.47	33.95
6000	0.37	387.07	7.08	394.52	0.35	128.96	7.72	137.03	0.58	69.07	8.18	77.83

Table 2. Simulation time in seconds for GPU and Multi-GPU codes, executing in the Phenom Workstation.

N	GPU				Multi-GPU				
	fill	evolve	output	total	fill	evolve	output	halo	total
500	0.07	0.16	0.05	0.29	0.25	0.44	0.07	1.64	2.38
1000	0.09	0.56	0.2	0.86	0.26	0.77	0.24	1.65	2.93
2000	0.12	1.82	0.78	2.72	0.32	1.53	0.97	1.72	4.54
4000	0.26	7.35	3.21	10.81	0.49	4	4.06	2.99	11.53
6000	0.48	18.55	8.11	27.15	0.74	9.9	9.3	2.65	22.6

of information between GPUs ("halo" table column). When the GPU is used in the Phenom workstation, four parallel processes are executed in the same GPU, which causes an increase in the communication time in the PCI-Express bus. Using four GPUs for the larger case shows an improvement in the simulation time, decreasing it from 26.4 to 19.3 seconds. Speedups can be seen in table 6.

Table 3. Simulation time in seconds for CPU serial, OpenMP and MPI codes, executing in the Mendieta Cluster.

N	CPU Serial				OpenMP				MPI			
	fill	evolve	output	total	fill	evolve	output	total	fill	evolve	output	total
500	0.03	2.21	0.04	2.25	0.005	1.30	0.06	1.37	0.006	0.41	0.1	0.52
1000	0.01	8.94	0.16	9.11	0.01	5.25	1.07	6.33	0.03	1.22	0.41	1.66
2000	0.04	40.27	0.62	40.93	0.04	25.33	1.05	26.41	0.14	4.35	1.54	6.03
4000	0.16	188.49	3.78	192.42	0.27	101.05	4.23	105.54	0.59	26.66	7.01	34.26
6000	0.34	376.71	5.58	382.64	0.35	225.56	9.66	235.57	1.32	59.59	13.86	74.76

Tests were also performed on the FX-4100 workstation, for a single simulation, with *N=2000* and *1000* steps. Results are: Serial = 44.42 s, OpenMP = 18.63 s, MPI = 9.82 s, GPU = 17.5 s, Multi-GPU = 16.94 s. The video card installed in this machine is a low-range card and, as a result, its performance compare to CPUs is far from what could be reached with a high-range card.

Table 4. Simulation time in seconds for GPU and Multi-GPU codes, executing in the Mendieta Cluster.

N	GPU				Multi-GPU				
	fill	evolve	output	total	fill	evolve	output	halo	total
500	3.78	0.13	1.15	5.06	3.99	0.51	0.19	0.43	5.11
1000	3.81	0.43	0.31	4.55	4.08	0.53	0.42	0.55	5.59
2000	3.9	1.64	0.8	6.34	4.06	0.75	1.73	0.8	7.34
4000	4.09	5.66	2.98	12.73	4.3	1.32	6.77	1.28	13.67
6000	4.21	14.36	13.35	31.92	4.7	3.43	15.05	3.21	26.39
6000	running in 2 GPU per node				1.48	3.29	14.31	0.18	19.27

Table 5. Speedups for all parallel codes vs the serial code in the Phenom workstation

dim	Speedup vs Serial Code			
	MPI	OMP	GPU	Multi-GPU
500	8.32	2.33	8	0.97
1000	7.44	2.8	11.92	3.49
2000	4.62	2.64	14.91	8.94
4000	5.01	2.84	15.72	14.73
6000	5.07	2.88	14.53	17.46
AVG	6.09	2.70	13.02	9.12

Table 6. Speedups for all parallel codes vs the serial code in the Mendieta cluster

dim	Speedup vs Serial Code			
	MPI	OMP	GPU	Multi-GPU
500	4.33	1.64	0.45	0.44
1000	5.50	1.44	2.00	1.63
2000	6.79	1.55	6.46	5.58
4000	5.62	1.82	15.11	14.08
6000	5.12	1.62	11.99	14.50
6000 in 2 GPUs per node				19.86
AVG	5.47	1.62	7.20	9.35

5 Conclusions and future works

The implementation of the popular Game of Life [4] was used in this paper as a possible introduction to the problem of parallel processing of a CA on CPU+GPU hybrid environments. Improvements in the execution times in the pure GPU implementation and the Multi-GPU implementation have been achieved, with speed-ups approaching 20 x, when compared to a serial CPU implementation. The MPI implementation of the code gave better timing than the implementation using OpenMP, but the development time for the MPI code was higher. The OpenMP implementation did not perform as expected, and it would be necessary to carry out a detail code analysis with a tool like perf [26] to improve it. Using the Multi-GPU code provides significant speed-ups, but only for large grids (above a few million grid points).

There are several possible improvements for the codes presented here. For instance, the codes could be adapted to support square grids with N being an odd number, non-square grids, and odd number of processes. In addition, MPI and Multi-GPU codes deal with neighbours in a simple way, but MPI provides functions which could be used to perform these tasks more efficiently. There was no optimization of the codes, beyond the standard compiler optimization. There are several ways to optimize and improve performance in GPU codes: management of shared memory, monitoring and reducing the number of registers

used by each kernel, etc. Starting with CUDA 4.0, GPUDirect (`http://goo.gl/g7D1p`) technology is available and supported by MVAPICH [27], and data can be directly transferred between GPUs without passing through the main memory system.

Once an efficient CA is implemented in multiple GPUs, one could move to related problems, such as the Reaction-Diffusion Equations [19] [20], the Cahn-Hilliard equation [17][18], or a CA representing some general problem of interest [28].

6 Acknowledgements

E. Millán acknowledges support from a CONICET doctoral scholarship. E.M. Bringa thanks support from a SeCTyP2011-2013 project and PICT2009-0092.

References

1. CUDA from NVIDIA: `http://www.nvidia.com/cuda`
2. OpenCL, The open standard for parallel programming of heterogeneous systems: `http://www.khronos.org/opencl/`
3. Dokulil, J., Bajrovic, E., Benkner, S., Pllana, S., Sandrieser, M. and Bachmayer, B. Efficient Hybrid Execution of C++ Applications using Intel (R) Xeon Phi (TM) Coprocessor. arXiv preprint arXiv:1211.5530 (2012).
4. Gardner, M. The fantastic combinations of John Conway's new solitaire game "life". Scientific American 223 (October 1970) pp 120-123.
5. Wolfram, S. Cellular Automata as models of complexity. Nature Vol. 311. pp 419-424. 1984.
6. Ganguly, N., Sikdar, B.K., Deutsch, A., Canright, G. and Chaudhuri, P.P. A survey on Cellular Automata. Centre for High Performance Computing, Dresden University of Technology. 2003.
7. Alt, W., Deutsch, A., and Dunn, G., editors. Dynamics of Cell and Tissue Motion. Birkhuser, Basel, 1997.
8. Packard, N.H. Lattice Models for Solidication and Aggregation. In First International
9. Schonsch, B. Propagation of Fronts in Cellular Automata. Physica D, 80:433450, October 2002.
10. Caux, J., Siregar, P., Hill, D.: Accelerating 3D Cellular Automata Computation with GP-GPU in the Context of Integrative Biology. In: Salcido, A. (ed.) Cellular Automata - Innovative Modelling for Science and Engineering. InTech. ISBN: 978-953-307-172-5. 2011.
11. Arvizu, C., Hector M., Trueba-Espinosa, A., and Ruiz-Castilla, J. Autmata Celular Estocstico paralelizado por GPU aplicado a la simulacin de enfermedades infecciosas en grandes poblaciones. Acta Universitaria 22.6: pp. 16-22. 2012.
12. Lopez-Torres, M. R., Guisado, J. L., Jimnez-Morales, F., & Diaz-del-Rio, F. GPU-based cellular automata simulations of laser dynamics. Jornadas Sarteco 2012.
13. Adamatzky, A., ed. Game of life cellular automata. Springer, 2010.
14. Alaniz, M., Bustos, F., Gil-Costa, V., Printista, M. Motor de simulacin para modelos de Automata Celular. 2nd International Symposium on Innovation and Technology ISIT 2011. 28-30 Noviembre, Lima Peru. Noviembre 2011. ISBN: 978-612-45917-2-3. pp. 66-71. 2011.

15. Rumpf, T. Conways Game of Life accelerated with OpenCL. In Proceedings of the Eleventh International Conference on Membrane Computing (CMC11), p. 459. book-on-demand. de, 2010.
16. Tlke, J. Implementation of a Lattice Boltzmann kernel using the Compute Unified Device Architecture developed by nVIDIA. Computing and Visualization in Science 13, no. 1: 29-39. 2010.
17. Cahn J.W. and Hilliard J.E. Free energy of a non-uniform system III. Nucleation in a two point compressible uid, J.Chem.Phys., vol. 31, pp. 688699, 1959.
18. Hawick, K. and Playne, D. Modelling and visualising the cahn-hilliard-cook equation. Tech. rep., Computer Science, Massey University, 2008. CSTN-049.
19. Smoller, J. Shock waves and reaction-diffusion equations. In Research supported by the US Air Force and National Science Foundation. New York and Heidelberg, Springer-Verlag (Grundlehren der Mathematischen Wissenschaften. Volume 258), 600 p. (Vol. 258). 1983.
20. Ready, A cross-platform implementation of various reaction-diffusion systems. https://code.google.com/p/reaction-diffusion/
21. The OpenMP API specification for parallel programming: `http://openmp.org/`
22. Newman, R.; MPI C version by Xianneng Shen. LAPLACE MPI Laplace's Equation in a Rectangle, Solved With MPI. `http://people.sc.fsu.edu/~jburkardt/c_src/laplace_mpi/laplace_mpi.html`
23. Lorna, S. and Bull, M. Development of mixed mode MPI/OpenMP applications. Scientific Programming 9, no. 2. pp 83-98. 2001.
24. Brown, W.M., Wang, P., Plimpton, S.J. and Tharrington, A.N. Implementing molecular dynamics on hybrid high performance computers short range forces. Computer Physics Communications 182. pp 898-911. 2011.
25. Millán, E.N., García Garino, C. and Bringa, E. Parallel execution of a parameter sweep for molecular dynamics simulations in a hybrid GPU/CPU environment. XVIII Congreso Argentino de Ciencias de la Computación 2012 (CACIC), Bahia Blanca, Buenos Aires. ISBN-978-987-1648-34-4. 2012.
26. perf: Linux profiling with performance counters. `https://perf.wiki.kernel.org/index.php/Main_Page`
27. MVAPICH: MPI over InfiniBand, 10GigE/iWARP and RoCE. `http://mvapich.cse.ohio-state.edu/`
28. Tissera, P., Printista, A. M., and Errecalde, M. L. Evacuation simulations using cellular automata. Journal of Computer Science & Technology, 7. 2007.

Evaluating tradeoff between recall and performance of GPU Permutation Index

Mariela Lopresti, Natalia Miranda, Mercedes Barrionuevo,
Fabiana Piccoli, Nora Reyes

LIDIC. Universidad Nacional de San Luis,
Ejército de los Andes 950 - 5700 - San Luis - Argentina
{omlopres, ncmiran, mdbarrio,mpiccoli, nreyes}@ unsl.edu.ar

Abstract. Query-by-content, by means of similarity search, is a fundamental operation for applications that deal with multimedia data. For this kind of query it is meaningless to look for elements exactly equal to a given one as query. Instead, we need to measure the dissimilarity between the query object and each database object. This search problem can be formalized with the concept of metric space. In this scenario, the search efficiency is understood as minimizing the number of distance calculations required to answer them. Building an index can be a solution, but with very large metric databases is not enough, it is also necessary to speed up the queries by using high performance computing, as GPU, and in some cases is reasonable to accept a fast answer although it was inexact. In this work we evaluate the tradeoff between the answer quality and time performance of our implementation of *Permutation Index*, on a pure GPU architecture, used to solve in parallel multiple approximate similarity searches on metric databases.

1 Introduction

Similarity search is a fundamental operation for applications that deal with multimedia data. For a query in a multimedia database it is meaningless to look for elements exactly equal to a given one as query. Instead, we need to measure the similarity (or dissimilarity) between the query object and each object of the database. The similarity search problem can be formally defined through the concept of metric space. The metric space model is a paradigm that allows to model all the similarity search problems. A metric space (X, d) is composed of a universe of valid objects X and a distance function defined among them, that determines the similarity (or dissimilarity) between two given objects and satisfies properties which make it a metric. Given a dataset of n objects, a query can be trivially answered by performing n distance evaluations, but sequential scan does not scale for large problems. The reduction of number of distance evaluations is important to achieve better results. Therefore, in many cases preprocessing the dataset is a good option to solve queries with as few distance computations as is possible. An index helps to retrieve the objects from the database that are relevant to the query by making much less than n distance evaluations during searches [1]. One of these indices is the *Permutation Index* [2].

The *Permutation Index* is a good representative of approximate similarity search algorithms to solve *inexact similarity searching* [3]. In this kind of similarity search, accuracy or determinism is traded for faster searches [1, 4]. Inexact similarity searching is reasonable in many applications because the metric-space

modelizations already involve an approximation to reality; hence, a second approximation at search time is usually acceptable.

Moreover, for very large metric database is not enough to preprocess the dataset by building an index, it is also necessary to speed up the queries by using high performance computing (HPC). In order to employ HPC to speedup the preprocess of the dataset to obtain an index, and to answer posed queries, the Graphics Processing Unit (GPU) represents a good alternative. The GPU is attractive in many application areas for its characteristics, especially because of its parallel execution capabilities and fast memory access. They promise more than an order of magnitude speedup over conventional processors for some non-graphics computations.

In metric spaces, the indexing and query resolution are the most common operations. They have several aspects that accept optimizations through the application of HPC techniques. There are many parallel solutions for some metric space operations implemented to GPU. Querying by k-Nearest Neighbors (k-NN) has concentrated the greatest attention of researchers in this area, so there are many solutions that consider GPU. In [5–9] differents proposal are made, all of them are improvements to brute force algorithm (sequential scan) to find the k-NN of a query object.

The goal of this work is to analyze the tradeoff betwen the quality of similarity queries answer and time performance, using a parallel permutation index implemented on GPU. In this analysis particularly we consider the known measures from information retrieval area for answer quality and we consider the achieved performance of our parallel implementation of *Pemutation Index*.

The paper is organized as follows: the next sections describe all the previous concepts. Sections 4 and 5 sketch the characteristics of our proposal and its empirical performance. Finally, the conclusions and future works are exposed.

2 Metric Space Model

A metric space (X, d) is composed of a universe of valid objects X and a distance function $d : X \times X \rightarrow R^+$ defined among them. The distance function determines the similarity (or dissimilarity) between two given objects and satisfies several properties such as strict positiveness (except $d(x, x) = 0$, which must always hold), symmetry ($d(x, y) = d(y, x)$), and the triangle inequality ($d(x, z) \leq d(x, y) + d(y, z)$). The finite subset $U \subseteq X$ with size $n = |U|$, is called the *database* and represents the set of objects of the search space. The distance is assumed to be expensive to compute, hence it is customary to define the search complexity as the number of distance evaluations performed, disregarding other components. There are two main queries of interest [1, 4]: Range Searching and the k-NN. The goal of a range search $(q, r)_d$ is to retrieve all the objects $x \in U$ within the radius r of the query q (i.e. $(q, r)_d = \{x \in U / d(q, x) \leq r\}$). In k-NN queries, the objective is to retrieve the set k-NN(q)$\subseteq U$ such that $\mid k\text{-NN}(q) \mid = k$ and $\forall x \in k\text{-NN}(q), v \in U \wedge v \notin k\text{-NN}(q), d(q, x) \leq d(q, v)$.

When an index is defined, it helps to retrieve the objects from U that are relevant to the query by making much less than n distance evaluations during searches. The saved information in the index can vary, some indices store a subset of distances between objects, others maintain just a range of distance values. In general, there is a tradeoff between the quantity of information maintained in the index and the query cost it achieves. As more information an index stores (more memory it uses), lower query cost it obtains. However, there are some indices

that use memory better than others. Therefore in a database of n objects, the most information an index could store is the $n(n-1)/2$ distances among all element pairs from the database. This is usually avoided because $O(n^2)$ space is unacceptable for realistic applications [10].

Proximity searching in metric spaces usually are solved in two stages: preprocessing and query time. During the preprocessing stage an index is built and it is used during query time to avoid some distance computations. Basically the state of the art in this area can be divided in two families [1]: *pivot based algorithms* and *compact partition based algorithms*.

There is an alternative to "exact" similarity searching called *approximate similarity searching* [3], where accuracy or determinism is traded for faster searches [1, 4], and encompasses *approximate* and *probabilistic algorithms*. The goal of approximate similarity search is to reduce *significantly* search times by allowing some errors in the query output. In these algorithms one usually has a threshold ϵ as parameter, so that the retrieved elements are guaranteed to have a distance to the query q at most $(1+\epsilon)$ times of what was asked for [11]. Probabilistic algorithms on the other hand state that the answer is correct with high probability. In approximate algorithms one usually has a threshold ϵ as parameter, so that the retrieved elements are guaranteed to have a distance to the query q at most $(1+\epsilon)$ times of what was asked for. This relaxation gives faster algorithms as the threshold ϵ increases [11, 12]. Probabilistic algorithms on the other hand state that the answer is correct with high probability [13, 14].

2.1 Quality Measures of Approximate Search

As it is aforementioned, an approximate similarity searching can obtain an inexact answer. That is, if a k-NN query of an element $q \in X$ is posed to the index, it answers with the k elements viewed as the k closest elements from U between only the elements that are actually compared with q. However, as we want to save as many distance calculations as we can, q will not be compared against many potentially relevant elements. If the exact answer of k-NN$(q) = \{x_1, x_2, \ldots, x_k\}$, it determines the radius $r_k = \max_{1 \leq i \leq k}\{d(x_i, q)\}$ needed to enclose these k closest elements to q. An approximate answer of k-NN(q) could obtain some elements z whose $d(q, z) > r_k$. For the other hand, an approximate range query of $(q, r)_d$ can answer a subset of the exact answer, because it is possible that the algorithm did not have reviewed all the relevant elements. However, all the answered elements will be at distance less or equal to r, so they belong to the exact answer to $(q, r)_d$.

In most of information retrieval (IR) systems it is necessary to evaluate retrieval effectiveness [15]. The judgements of document relevance used to evaluate effectiveness have some problems of subjectivity and unrealiableness. That is, different judges will assign different relevance values to a document retrieved in response to a given query. The seriousness of the problem is the subject of debate, with many IR researchers arguing that the relevance judgment reliability problem is not sufficient to invalidate the experiments that use relevance judgments. Many measures of retrieval effectiveness have been proposed. The most commonly used are *recall* and *precision*.

Recall is the ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database. *Precision* is the ratio of the number of relevant documents retrieved over the total number of documents retrieved. Both recall and precision take on values between 0 and 1.

Since one often wishes to compare IR performance in terms of both recall and precision, methods for evaluating them simultaneously have been developed

In general IR systems, only in small test collections, the denominator of both ratios is generally unknown and must be estimated by sampling or some other method. However, in our case we can obtain the exact answer for each query q, as the set of relevant elements for this query in U.

By this way it is possible to evaluate both measures for an approximate similarity search index. For each query element q the exact k-NN$(q) = Rel(q)$ is determined with some exact metric access method. The approximate-k-NN$(q) = Retr(q)$ is answered with an approximate similarity search index, let be the set $Retr(q) = \{y_1, y_2, \ldots, y_k\}$. It can be noticed that the approximate search will also return k elements, so $|Retr(q)| = |Rel(q)| = k$. Thus, we can determine the number of the k elements obtained which are relevant to q by verifying if $d(q, y_i) \leq r_k$; that is $|Rel(q) \cap Retr(q)|$. In this case both measures are coincident: recall $= \frac{|Rel(q) \cap Retr(q)|}{|Rel(q)|} = \frac{|Rel(q) \cap Retr(q)|}{k}$ and precision $= \frac{|Rel(q) \cap Retr(q)|}{|Retr(q)|} = \frac{|Rel(q) \cap Retr(q)|}{k}$, and will allow us to evaluate the effectiveness of our proposal. In range queries the precision measure is always equal to 1. Thus, we decide to use recall in order to analyze the retrieval effectiveness of our proposal, both in k-NN and range queries.

2.2 GPGPU

Mapping general-purpose computation onto GPU implies to use the graphics hardware to solve any applications, not necessarily of graphic nature. This is called GPGPU (General-Purpose GPU), GPU computational power is used to solve general-purpose problems [16]. The parallel programming over GPUs has many differences from parallel programming in typical parallel computer, the most relevant are: *The number of processing units*, *CPU-GPU memory structure* and *Number of parallel threads*.

Every GPGPU program has many basic steps, first the input data transfers to the graphics card. Once the data are in place on the card, many threads can be started (with little overhead). Each thread works over its data and, at the end of the computation, the results should be copied back to the host main memory. Not all kind of problem can be solved in the GPU architecture, the most suitable problems are those that can be implemented with stream processing and using limited memory, i.e. applications with abundant parallelism.

The Compute Unified Device Architecture (CUDA), supported from the NVIDIA Geforce 8 Series, enables to use GPU as a highly parallel computer for non-graphics applications [16, 17]. CUDA provides an essential high-level development environment with standard C/C++ language. It defines the GPU architecture as a programmable graphic unit which acts as a coprocessor for CPU. It has multiple streaming multiprocessors (SMs), each of them contains several (eight, thirty-two or forty-eight, depending GPU architecture) scalar processors (SPs). The CUDA programming model has two main characteristics: the parallel work through concurrent threads and the memory hierarchy. The user supplies a single source program encompassing both host (CPU) and *kernel* (GPU) code. Each CUDA program consists of multiple phases that are executed on either CPU or GPU. All phases that exhibit little or no data parallelism are implemented in CPU. Contrary, if the phases present much data parallelism, they are coded as *kernel* functions in GPU. A *kernel* function defines the code to be executed by each thread launched in a parallel phase.

3 Sequential Permutation Index

Let $\mathcal{P}$ be a subset of the database U, $\mathcal{P} = \{p_1, p_2, \ldots, p_m\} \subseteq U$, that is called the permutants set. Every element x of the database sorts all the permutants according to the distances to them, thus forming a permutation of $\mathcal{P}$: $\Pi_x = \langle p_{i_1}, p_{i_2}, \ldots p_{i_m} \rangle$. More formally, for an element $x \in U$, its permutation Π_x of $\mathcal{P}$ satisfies $d(x,\Pi_x(i)) \leq d(x,\Pi_x(i+1))$, where the elements at the same distance are taken in arbitrary, but consistent, order. We use $\Pi_x^{-1}(p_{i_j})$ for the *rank* of an element p_{i_j} in the permutation Π_x. If two elements are similar, they will have a similar permutation [2].

Basically, the permutation based algorithm is an example of probabilistic algorithm, it is used to predict proximity between elements, by using their permutations. The algorithm is very simple: In the offline preprocessing stage it is computed the permutation for each element in the database. All these permutations are stored and they form the index. When a query q arrives, its permutation Π_q is computed. Then, the elements in the database are sorted in increasing order of a similarity measurement between permutations, and next they are compared against the query q following this order, until some stopping criterion is achieved. The similarity between two permutations can be measured, for example, by *Kendall Tau*, *Spearman Rho*, or *Spearman Footrule* metrics [18]. All of them are metrics, because they satisfy the aforementioned properties. We use the Spearman Footrule metric because it is not expensive to compute and according to the authors in [2] it has a good performance to predict proximity between elements. The Spearman Footrule distance is the *Manhattan distance* L_1, that belongs to the Minkowsky's distances family, between two permutations. Formally, Spearman Footrule metric F is defined as: $F(\Pi_x, \Pi_q) = \sum_{i=1}^{m} |\Pi_x^{-1}(p_i) - \Pi_q^{-1}(p_i)|$.

At query time we first compute the real distances $d(q, p_i)$ for every $p_i \in \mathcal{P}$, then we obtain the permutation Π_q, and next we sort the elements $x \in U$ into increasing order according to $F(\Pi_x, \Pi_q)$ (the sorting can be done incrementally, because only some of the first elements are actually needed). Then U is traversed in that sorted order, evaluating the distance $d(q, x)$ for each $x \in U$. For range queries, with radius r, each x that satisfies $d(q, x) \leq r$ is reported, and for k-NN queries the set of the k smallest distances so far, and the corresponding elements, are maintained. The database traversal is stopped at some point f, and the rest of the database elements are just ignored. This makes the algorithm probabilistic, as even if $F(\Pi_q, \Pi_x) < F(\Pi_q, \Pi_v)$ it does not guarantee that $d(q, x) < d(q, v)$, and the stopping criterion may halt the search prematurely. On the other hand, if the order induced by $F(\Pi_q, \Pi_x)$ is close to the order induced by the real distances $d(q, u)$, the algorithm performs very well. The efficiency and the quality of the answer obviously depend on f. In [2], the authors discuss a way to obtain good values for f for sequential processing.

4 GPU-Permutation Index

The Figure 1 shows the GPU-CUDA system to work with a permutation index: the processes of indexing and querying. The Indexing process has two stages and the Querying process four steps. In this last process, we pay special attention to one step: the sorting.

Building a permutation index in GPU involves at least two steps. The first step (*Distance(O,P)*) calculates the distance among every object in database and

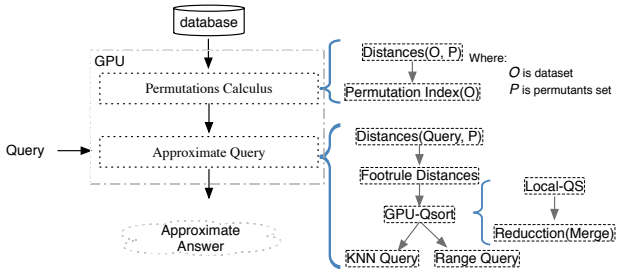


Fig. 1. Indexing and Querying in GPU-CUDA permutation index.

the permutants. The second one (*Permutation Index(O)*) sets up the signatures of all objects in database, i.e. all object permutations. The process input is the database and the permutants. At process end, the index is ready to be queried. The idea is to divide the work in threads blocks, each thread calculates the object permutation according to a global permutants set.

In $Distances(O, P)$, the number of blocks will be defined according of the size of the database and the number of threads per block which depends of the quantity of resources required by each block. At the end, each threads block save in device memory its calculated distances. This stage requires a structure of size $m \times n$ (m: permutants number and n: database size) and an auxiliar structure in the shared memory of block (It stores the permutants, if the permutants size is greater than auxiliar structure size, the process is repeated). The second step ($Permutation\ Index(O)$) takes all calculated distances in the previous step and determines the permutations of each object in database: its signature. To stablish the object permutation, each thread considers an object in database and sorts the permutants according to their distance. The output of second step is the *Permutation Index*, which is saved in device memory. Its size is $n \times m$.

The pemutation index allows to answer to all kinds of queries in approximated manner. Queries can be *"by range"* or *"*k-NN". This process implies four steps. In the first, the permutation of query object is computed. This task is carried out by so many threads as permutants exist. The next step is to contrast all permutations in the index with query permutation. Comparison is done through the *Footrule* distance, one thread by object in database. In the third step, it sorts the calculated *Footrule* distances. Finally, depending of query kind, the selected objects have to be evaluated. In this evaluation, the *Euclidean distance* between query object and each candidate element is calculated again. Only a database percentage is considered for this step, for example the 10% (it can be a parameter). If the query is by range, the elements in the answer will be those that their distances are less than reference range. If it is k-NN query, once each thread computes the *Euclidean distance*, all distances are sorted and the results are the first k elements of sorted list.

As sorting methodology, we implement the Quick-sort in the GPU, GPU-Qsort. The designed algorithm takes into account the highly parallel nature of graphics processors (GPUs) and the CUDA capabilities 1.2 or higher. Its main characteristics are: iterative algorithm and heavy use of shared memory of each block, you can find an detailed description in[19].

In large-scale systems such as Web Search Engines indexing multimedia content, it is critical to deal efficiently with streams of queries rather than with

single queries. Therefore, it is not enough to speed up the time to answer only one query, but it is necessary to leverage the capabilities of the GPU to parallely answer several queries. So we have to show how to achieve efficient and scalable performance in this context. We need to devise algorithms and optimizations specially tailored to support high-performance parallel query processing in GPU. GPU has characteristics of software and hardware which allow us to think in to solve many approximated queries in parallel. The represented system in Figure 1 is modified and it is shown in Figure 2. In this, it can be observed that the permutation index is built once and then is used to answer many queries.

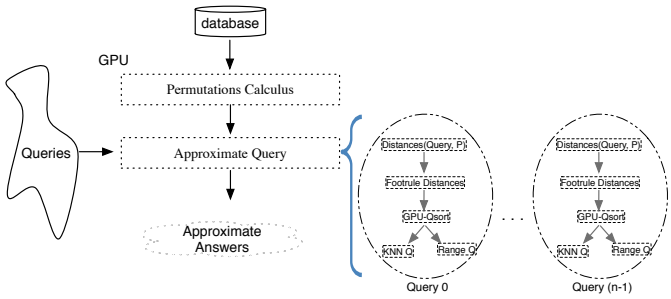


Fig. 2. Solving many queries in GPU-CUDA permutation index.

In order to answer parallely many approximate queries, GPU receives the queries set and it has to solve all of them. Each query, in parallel, applies the process explained in 1, therefore the number of needed resources for this is equal to the resources amount to compute one query multiplied the number of queries solved in parallel. The number of queries to solve in parallel is determined according to the GPU resources mainly its memory. If Q are parallel queries, m the needed memory quantity per query and i the needed memory by permutation index, $Q * m + i$ is the total required memory to solve Q queries in parallel. Once the Q parallel queries are solved, the results are sent from the GPU to the CPU through a single transfer via PCI-Express.

Solving many queries in parallel involves carefully manage the blocks and their threads. At the same time, blocks of different queries are accessed in parallel. Hence it is important a good administration of threads: which query it is solved and which database element it is responsible. The task is possible by establishing a relationship among *Thread Id*, *Block Id*, *Query Id*, and *Database Element*.

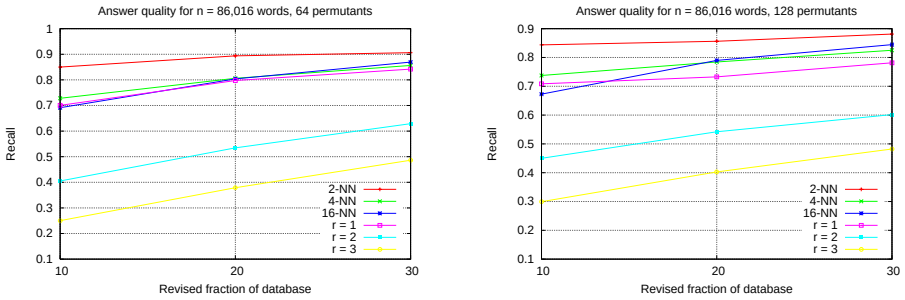
5 Experimental Results

Our experiments considered a metric database of 86,016 English words and using the *Levenshtein* distance, also called *edit* distance [20]. The analysis was made for a GeForce GPU GTX520MX whose characteristics (Global Memory: 1024MB, SM:1, SP:48, Clock rate:1.8GHz, Compute Capability: 2.1). The CPU is an Intel core i3, 2.13 GHz and 3 GB of memory.

The experiments consider for k-NN searches the values of k: 2, 4, and 16; and for range the radii: 1, 2, and 3. For the parameter f of the Permutation Index,

that indicates the fraction of database revised during searches, we consider 10%, 20%, and 30% of the database size. The number of permutants used for the index are 64 and 128. In each case the results shown are the average over 1000 different queries and 80 solved queries in parallel. In this paper, we do not display the speed up of construction of *Permutation Index*. These results are illustrated in [21].

Our focus is to evaluate the tradeoff between the answer quality and time performance of our parallel index with respect to the sequential index. For each k-NN or range query we have previously obtained the exact answer, that is $Rel()$, and we obtain the approximate answer $Retr()$. Figure 3 illustrates the average quality answer obtained for both kinds of queries, considering the Permutation Index respectively with 64 (Figure 3(a)) and 128 (Figure 3(b)) permutants. As it can be noticed, the Permutation Index retrieves a high percentage of exact answer only reviewing a little fraction of the database. For example, the 10% retrieves 85% for 2-NN queries both with 64 and 128 permutants. It needs to review the 20% to retrieve almost 80% of exact answer for $k = 4$ and $k = 16$ with 64 and 128 permutants. The effectiveness in range queries decreases as the radius grows. For $r = 1$ the index retrieves almost 80% of the relevant objects.

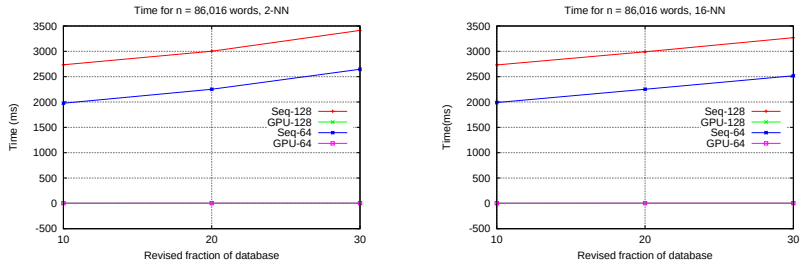


(a) 64 permutants (b) 128 permutants

Fig. 3. Recall of approximate-k-NN and range queries obtained with Permutation Index.

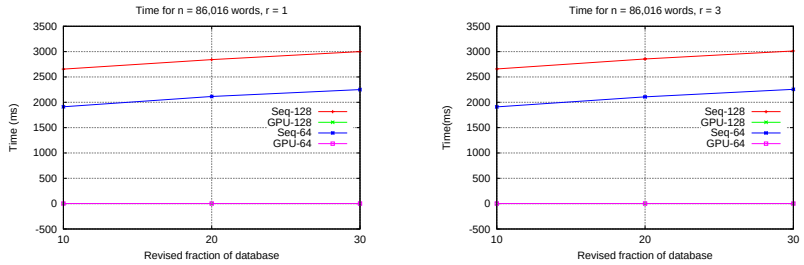
Figures 4 and 5 show the obtained times by k-NN and range queries for three f values and all the number of permutants considered. In these results, 80 queries are solved in parallel on GPU. As it can be noticed, the parallel times are so smaller than the corresponding sequential times. In both types of queries the achieved speed up is very good, it can be observed the same behavior for all options of our parallel solution, they are independent of the number of permutants and fraction f of database to be revised.

For lack of space, despite of we have tested another database sizes, we show only the results for 86,016 elements, but the other sizes have yielded similar results.



(a) 2-NN (b) 16-NN

Fig. 4. Time of k-NN queries obtained with Sequential and Parallel Permutation Index.



(a) Range query r = 1 (b) Range query, r = 3

Fig. 5. Time of Range queries obtained with Sequential and Parallel Permutation Index.

6 Conclusions

When we work with databases, there are different realities where it is not enough to speed up the time to answer only one query, but it is necessary to solve several queries at the same time. In this work we present a solution to solve many queries in parallel taking advantage of GPU architecture: it is a massively parallel architecture, it has a high throughput because its capacity of parallel processing for thousands of threads, and verify the correctness of obtained results.

The implemented GPU-*Pemutation Index* showed a good performance, allowing us to increase the fraction f of database that will be examined to obtain better and accurate approximate results. This affirmation is made in function of an extensive validation process carried out to guarantee the quality of the solution provided by the GPU.

As future task, we need to validate every performance parameters: recall, speed up and throughput, with other kinds of database, comparing with other solutions that apply GPU in the scenario of metric space approximate searches.

References

1. E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquín, "Searching in metric spaces," *ACM Comput. Surv.*, vol. 33, no. 3, pp. 273–321, 2001.

2. E. Chavez, K. Figueroa, and G. Navarro, "Effective proximity retrieval by ordering permutations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1647–1658, 2008.
3. P. Ciaccia and M. Patella, "Approximate and probabilistic methods," *SIGSPATIAL Special*, vol. 2, no. 2, pp. 16–19, Jul. 2010.
4. P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search: The Metric Space Approach*, ser. Advances in Database Systems, vol.32. Springer, 2006.
5. R. J. Barrientos, J. Gomez, C. Tenllado, M. Prieto, and M. Marin, "knn query processing in metric spaces using gpus," in *17th International European Conference on Parallel and Distributed Computing*, L. N. i. C. S. Springer, Ed., vol. 6852, 2011, pp. 380–392.
6. V. Garcia, E. Debreuve, F. Nielsen, and M. Barlaud, "k-nearest neighbor search: fast GPU-based implementations and application to high-dimensional feature matching," in *IEEE Intern. Conf. on Image Processing*, Hong Kong, Sept. 2010.
7. K. Kato and T. Hosino, "Solving k-nearest neighbor problem on multiple graphics processors," in *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, CCGRID*, ACM, Ed., 2010, pp. 769–773.
8. S. Liang, Y. Liu, C. Wang, and L. Jian, "Design and evaluation of a parallel k-nearest neighbor algorithm on CUDA-enabled GPU," in *IEEE 2nd Symposium on Web Society (SWS)*, 2010, pp. 53 – 60.
9. R. Uribe, P. Valero, E. Arias, J. L. Sánchez, and D. Cazorla, "A GPU-Based Implementation for Range Queries on Spaghettis Data Structure," in *ICCSA (1)*, ser. Lecture Notes in Computer Science, vol. 6782. Springer, 2011, pp. 615–629.
10. K. Figueroa, E. Chávez, G. Navarro, and R. Paredes, "Speeding up spatial approximation search in metric spaces," *ACM Journal of Experimental Algorithmics*, vol. 14, p. article 3.6, 2009.
11. B. Benjamin and G. Navarro, "Probabilistic proximity searching algorithms based on compact partitions," *Discrete Algorithms*, vol. 2, no. 1, pp. 115–134, Mar. 2004.
12. K. Tokoro, K. Yamaguchi, and S. Masuda, "Improvements of tlaesa nearest neighbour search algorithm and extension to approximation search," in *Proceedings of the 29th Australasian Computer Science Conference - Volume 48*, ser. ACSC '06. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2006, pp. 77–83.
13. A. Singh, H. Ferhatosmanoglu, and A. Tosun, "High dimensional reverse nearest neighbor queries," in *The 12th intern. conf. on Information and knowledge management*, ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 91–98.
14. F. Moreno, L. Mic, and J. Oncina, "A modification of the laesa algorithm for approximated k-nn classification," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 47 – 53, 2003.
15. R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.
16. D. B. Kirk and W. W. Hwu, *Programming Massively Parallel Processors, A Hands on Approach*. Elsevier, Morgan Kaufmann, 2010.
17. NVIDIA, "Nvidia cuda compute unified device architecture, programming guide version 4.2." in *NVIDIA*, 2012.
18. R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *Proc. of the 40th annual ACM-SIAM symposium on Discrete algorithms, SODA '03*. Philadelphia, USA: Society for Industrial and Applied Mathematics, 2003, pp. 28–36.
19. M. Lopresti, N. Miranda, F. Piccoli, and N. Reyes, "Permutation index and gpu to solve efficiently many queries," in *VI Latin American Symposium on High Performance Computing, HPCLatAm 2013*, 2013, pp. 101–112.
20. V. I. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones," *Problems of Information Transmission*, vol. 1, pp. 8–17, 1965.
21. M. Lopresti, N. Miranda, F. Piccoli, and N. Reyes, "Efficient similarity search on multimedia databases," in *XVIII Congreso Argentino de Ciencias de la Computacin, CACIC 2012*, 2012, pp. 1079–1088.

Managing Receiver-Based Message Logging Overheads in Parallel Applications

Hugo Meyer*, Dolores Rexachs, and Emilio Luque

Computer Architecture and Operating Systems Department,
University Autonoma of Barcelona, Barcelona, Spain
{hugo.meyer}@caos.uab.es
{dolores.rexachs,emilio.luque}@uab.es
http://www.uab.es

Abstract. Using rollback-recovery based fault tolerance (FT) techniques in applications executed on Multicore Clusters is still a challenge, because the overheads added depend on the applications' behavior and resource utilization. Many FT mechanisms have been developed in recent years, but analysis is lacking concerning how parallel applications are affected when applying such mechanisms. In this work we address the combination of process mapping and FT tasks mapping on multicore environments. Our main goal is to determine the configuration of a pessimistic receiver-based message logging approach which generates the least disturbance to the parallel application. We propose to characterize the parallel application in combination with the message logging approach in order to determine the most significant aspects of the application such as computation-communication ratio and then, according to the values obtained, we suggest a configuration that can minimize the added overhead for each specific scenario. In this work we show that in some situations is better to save some resources for the FT tasks in order to lower the disturbance in parallel executions and also to save memory for these FT tasks. Initial results have demonstrated that when saving resources for the FT tasks we can achieve 25% overhead reduction when using a pessimistic message logging aproach as FT support.

Keywords: Fault Tolerance, Mapping, Message Logging, Multicore, Overheads.

1 Introduction

Current High Performance Computing (HPC) systems are composed of nodes containing many processing units in order to execute more work in a short amount of time [1]. In order to take full advantage of the parallel environment,

* This research has been supported by the MICINN Spain under contract TIN2007-64974, the MINECO (MICINN) Spain under contract TIN2011-24384, the European ITEA2 project H4H, No 09011 and the Avanza Competitividad I+D+I program under contract TSI-020400-2010-120.

a good process mapping is essential. It is also important to consider that when executing parallel applications the fundamental objectives are: speedup as close as possible to the ideal (scalability) and efficient resource utilization.

Considering that applications are mapped into parallel environments in order to fulfill the above mentioned objectives, any disturbance may render all the mapping work useless. Currently, it is increasingly relevant to consider node failure probability since the mean time between failure in computer clusters has become lower [2] and this may cause loss of significant computation time in long-running applications. Indeed, successful completion of executions should be added to the list of fundamental objectives. In this vein FT techniques are gaining importance when running parallel applications. Nevertheless, FT mechanisms introduce disturbance to parallel applications in the form of overheads, which if not managed can result in large performance degradations, thus FT mechanisms that do no endanger scalability (uncoordinated approaches) are preferred.

Many recent works focus on finding the best checkpoint interval, or determining the best checkpoint or message logging approach for parallel applications [3][4] but few works address assigning resources for fault tolerance tasks considering applications' behavior [5].

When single-core clusters were the only option to execute parallel applications, there was not too many choices when talking about sharing resources. As there was only one computational core available, parallel applications share this resource (as well as the memory and cache levels) with the FT tasks if there was not dedicated resources. Considering that current clusters have more cores and memory levels, there is a need to develop mapping policies that allow parallel applications to coexist with the FT tasks in order to reduce the disturbance caused by these tasks. There is also important to consider that the number of cores has been multiplied by 8, 16, 32, 64 and usually the networks used in these clusters have not increase their speed to the same extent.

The main objective of this work is to determine the configuration of parallel applications in combination with a pessimistic receiver-based logging approach that minimizes the added overhead. We analyze parallel applications and obtain information that allows us to configure properly the FT tasks, specifically we determine if the best option is to share (compete for) resources with application processes or save resources for the FT tasks in order to reduce the introduced disturbance. In order to provide the configurations we consider the balance between computation and communication, message sizes and per-process memory consumption among other values.

The rest of the paper is organized as follows: Section 2 describes related work. Section 3 presents an analysis of the possible scenarios when executing a parallel application. Section 4 describes how to analyze a parallel application in order to find the most suitable message logging configuration. Section 5 shows the experimental validation and finally section 6 draws the main conclusions and mentions future works.

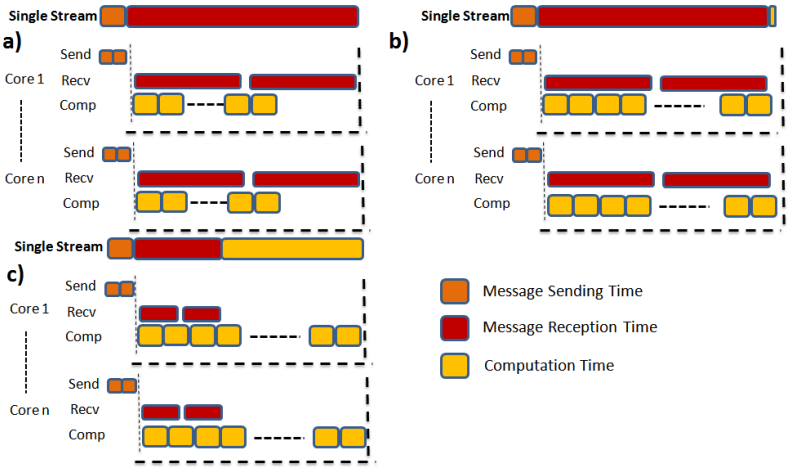


Fig. 1. Parallel Executions Scenarios in a SPMD App. a)Communication Bound. b) Computation and Communication overlapped. c) Computation Bound.

2 Related Work

In order to provide fault tolerance to parallel applications many strategies have been designed using message logging approaches [2][4][3]. Message logging approaches can sustain a much more adverse failure pattern, mainly due to a faster failure recovery. The main disadvantage of message log schemes is that they suffer from high overhead during failure-free executions [6], but they are an scalable solution since only failed processes must rollback, unless the domino effect is not addressed. Usually message logging techniques are used in combination with uncoordinated checkpoint approaches. Uncoordinated approaches are a good solution because there is not need for coordination between processes and there is no dependency on global components that could cause bottlenecks and compromise applications' scalability.

Following these lines, to develop this work we have used the RADIC (Redundant Array of Distributed and Independent Controllers) architecture [7], which uses a pessimistic receiver-based message logging technique in combination with an uncoordinated checkpoint approach in order to give application-transparent and scalable FT support for message passing applications.

In [3] a comparison between a pessimistic and optimistic sender-based logging approaches is presented where both seem to have a comparable performance. Nevertheless, when using sender-based approaches we should consider that in the presence of failures, processes that were not involved in the failure may need to re-send messages to restarted processes, and also garbage collection is complex. The pessimistic receiver based message logging approach of RADIC may be more costly than a sender based approach, but it guarantees that only failed processes will rollback to a previous state, without needing the intervention of other processes during re-execution.

In [8] was proposed a mechanism to reduce the overhead added using the pessimistic receiver-based message logging of RADIC. The technique consists in dividing messages into smaller pieces, so receptors can overlap receiving pieces

with the message logging mechanism. This technique and all the RADIC Architecture has been introduced into Open MPI in order to support message passing applications.

In [9] was presented an algorithm for distributing processes of parallel applications across processing resources paying attention to on-node hardware topologies and memory locales. When it comes to combine the mapping of FT tasks, specifically message logging tasks, with application process mapping, to date, no works have been published to the best of our knowledge.

3 Analyzing Parallel Applications Behavior

Current HPC parallel applications are executed on multicore systems, and the executions usually aim for almost lineal speedup and efficient resource utilization. In Figure 1 we present the three main scenarios possible when mapping applications in multicore systems. It is important to highlight that in this figure we are considering one iteration of a SPMD application. In Figure 1 we decompose the Single Stream in communications and computations operations. The main scenarios are:

1. **Communication bound**: Applications in which the processes are waiting because the communications take more time than the computations belong to this scenario. In Figure 1a we show how a communication bound application behaves (we are using as an example a SPMD application, where all processes do the same thing and each message goes from one process to another in a different core). In this figure we focus on showing how reception times (non-blocking send operations do not delay considerably the execution) can influence highly the execution time of a parallel application.
2. **Balanced Application**: This scenario is the best regarding efficient resource utilization, because the computational elements are working while the communication takes place. However, this behavior is very difficult to obtain because a previous study of the application is needed in order to completely overlap computations and communications (Figure 1b).
3. **Computation Bound**: When operators try to make a good use of the parallel environment they try to maintain the CPU efficiency high. Then in order to avoid the communication bound scenario it is recommended to give more workload per process which usually leads to a computation bound scenario. Figure 1c illustrates this scenario.

When characterizing a parallel application, it is also important to consider the number of processes that will be used, the number of nodes and the memory consumption of each process. This analysis should be done in combination with the analysis of the parallel environment in order to determine resource utilization. In this paper, we have characterized the parallel environment using application kernels and we consider the application phases (repetitive pieces of the parallel execution) that have the biggest weights during application execution.

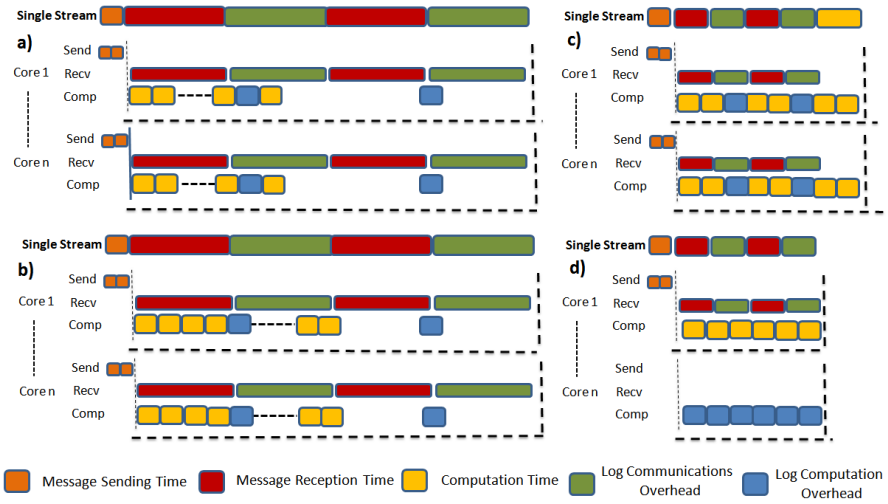


Fig. 2. Parallel Executions Scenarios with Message Logging. a)A communication bound application. b) A balanced application becomes comm. bound. c) A computation bound application stays as it was. d) A computation bound application becomes balanced.

In order to find the most appropriate configuration of the message logging approach, we should analyze how the parallel application and the logging approach coexist in the parallel machine. There will be two parallel applications that will compete for resources, thus it is critical to analyze the influence of FT in application behavior.

4 Analyzing Message Logging Processes Mapping

Most of the impact of a pessimistic receiver-based message logging protocol concentrates on communications and storage (memory or hard disks), but there is also a small impact on computations because FT tasks also need some CPU cycles in order to carry on their work.

For the analysis in this paper we have considered the pessimistic receiver-based message logging protocol used in RADIC. RADIC main components are shown on Figure 3, Protectors' main functions during the protection stage are: establish a heartbeat/watchdog mechanism between nodes (low CPU consumption operation, do not depend on application behavior) and to receive and manage message logs (CPU consumption depends on application) and checkpoints from observers (infrequent operation). All communications between processes go through Observers and each received message is sent to a corresponding logger thread (usually the protector of RADIC is drawn as an equilateral triangle, but in this case we have split it two right triangles to distinguish the main operations).

In order to reduce the impact of the pessimistic receiver-based logging protocol of RADIC we propose to save computational cores for the logger threads (Namely, threads that are in charge of receiving and logging messages of other processes), thus avoiding the competition for CPU between application processes

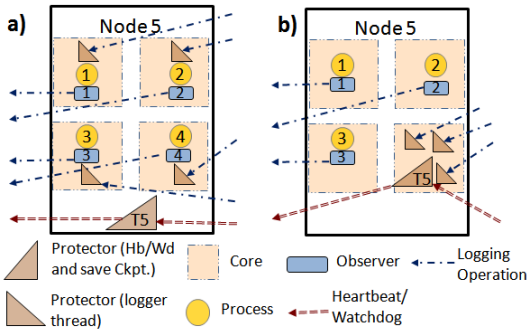


Fig. 3. RADIC Processes Mapping. a)Logging threads equitably distributed among cores. b) Protector with own resources.

and logger threads. According to this protocol every message should be logged in a different computing node (there are no dedicated nodes, but the usage of Spare Nodes is considered in [7]), then there is a considerable increase in the transmission time of each message when RADIC protection is activated. Thus, when executing a parallel application with RADIC the previous scenarios change (Figure 1), because the processes will be waiting a longer time for each message and the computation will be affected by the logger threads.

In order to reduce the overheads of the message logging approach, we analyzed how the application will be affected when introducing this logging technique. In Figure 2a we can observe how in a communication bound application, the difference between communications and computations becomes higher, and the overhead added in computations does not affect the iteration time. A balanced application without message logging will become communication bound when message logging is used (Figure 2b).

In these two scenarios the message logging overheads cannot be hidden, but when it comes to computation bound applications we can manage the mapping of the logger threads so as to distribute the overheads equally among the processes (Figure 2c). Alternatively, we can choose to save some computational cores in each node in order to avoid context switch between application processes and logger threads (Figure 2d).

Considering that many parallel applications are executed with balanced per-core workload, our default proposal is to distribute the overhead in computation produced by the logger threads among application processes residing in a node (Figure 3a). Moreover, we characterize the parallel application in order to find the computation-communication ratio, and if the application is computation bound, we analyze the overheads produced in computations. If these overheads make the application behave as in Figure 2c, we propose saving cores in each node in order to assign them to the logger threads, obtaining the behavior showed in Figure 2d. Figure 3b shows how we assign the logger threads and other proctectors' functionalities when using own resources for them.

As saving cores may make the initial mapping change, we also analyze if the new mapping does not negatively affect the execution, resulting in a worse performance than the default option.

Another important aspect that we analyze is the per-process memory consumption. This is significant because we have the option of storing the message log in memory instead of hard disks as this allows us to avoid bigger delays when storing messages. When we put less processes per node, we can save more memory for the message log, thus there is more time to overlap the flush-to-disk operation with receptions of new messages to log. Also, we can use longer checkpoint intervals if we consider an event-triggered checkpoint mechanism where a full message log buffer triggers a checkpoint.

5 Experimental Validation

The main approach presented in this paper focus on resource assignation to decrease logging overheads and save memory for FT tasks. In this section we present experimental evaluation that has been carried out in order to probe our hypothesis.

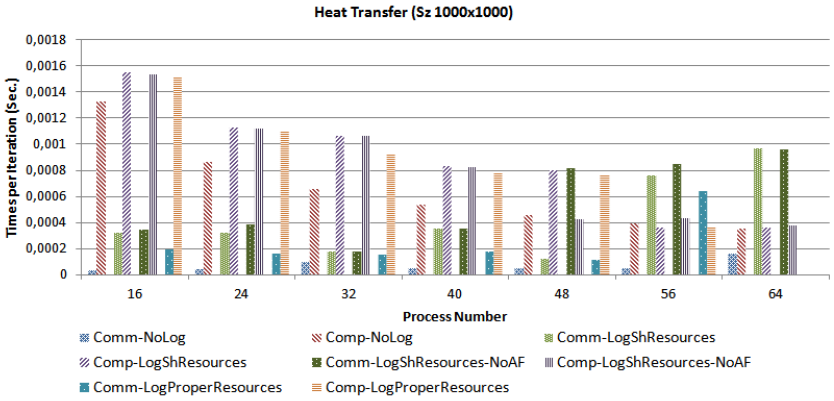
The experiments and characterizations have been made using a Dell PowerEdge M600 with 8 nodes, each node with 2 quad-core Intel® Xeon® E5430 running at 2.66 GHz. Each node has 16 GB of main memory and a dual embedded Broadcom® NetXtreme IITM 5708 Gigabit Ethernet. RADIC features have been integrated into Open MPI 1.7.

Most of the overhead added by a logging protocol affects communications. In order to lower the impact of a message logging technique we can assign more work per process which allow us to hide the overheads in communications (Figure 2c). However, if there are no available computational resources for the fault tolerance tasks, the overheads in computations could become relevant. Moreover, if we are executing a parallel application where memory consumption per process is high, there will be no room for the FT mechanisms.

When executing a parallel application with FT support is desirable to store checkpoints and message logs in main memory avoiding the file system, thus allowing FT mechanisms to execute faster. Also, if we consider an event triggered checkpoint mechanism where checkpoints take place when a message-log-buffer in memory is full and we save memory by executing less application processes per node we can use a bigger message-log-buffer, thus the checkpoint interval could be bigger.

Our testbed here is composed by two SPMD applications: a Heat Transfer application and a Laplace Solver. Both applications allow overlapping between the computation steps and the communication steps as was shown in Figure 2 and are written using non-blocking communications. The computation and communication times per iteration showed in bars in Figure 4 and Figure 5 are obtained by executing a few iterations of the parallel applications observing all processes and then selecting the slowest one for each number of processes. The execution times have been obtained using 10000 iterations.

In this experiments we have only considered the overlapped times (communication and computations) because they represent the higher portion of both applications. We have discarded the delays caused by desynchronization and



Overheads With Barriers			
# Process	Ovh-ShResources-NoAF	Ovh-ShResources	Ovh-ProperResources
16	55,25	51,75	46,55
24	69,42	61,96	56,94
32	44,06	41,24	28,94
40	116,03	86,14	71,52
48	123,38	96,99	75,58
56	129,86	99,10	75,40
64	54,58	43,28	

Overheads Without Barriers			
# Process	Ovh-ShResources-NoAF	Ovh-ShResources	Ovh-ProperResources
16	35,85	35,00	36,57
24	56,07	48,34	45,47
32	69,48	59,40	53,98
40	85,41	78,13	66,55
48	93,09	78,62	65,14
56	103,10	91,09	65,97
64	180,31	145,58	

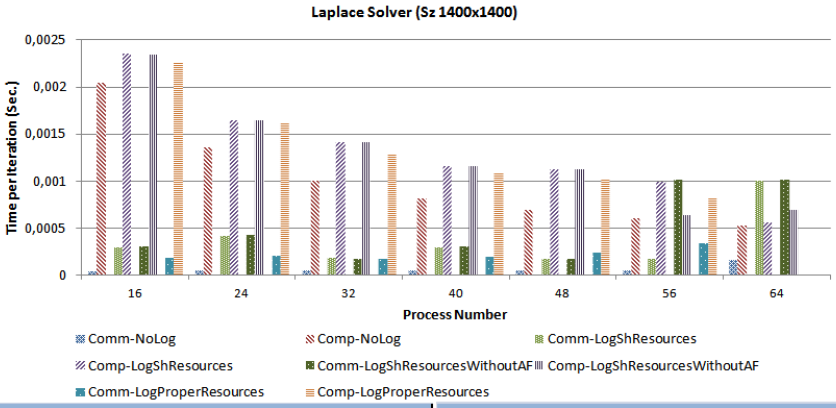
Fig. 4. Characterization results and Overhead Analysis of the Heat Transfer Application.

the computation time spent in computing the edges of each sub-matrix before sending it to the neighbors.

For both applications we have measure communication and computation times with the following options:

1. Without using message logging (Comm-NoLog and Comp-NoLog).
2. With message logging using all available cores in each node and giving affinity to each logger thread in order to ensure an equally distributed overhead in computation among all application processes (Comm-LogShResources and Comp-LogShResources).
3. With message logging using all available cores in each node without giving affinity to each logger thread (Comm-LogShResources-NoAF and Comp-LogShResources-NoAF).
4. With message logging saving one core per node and assigning all logger threads to the core not used by application processes (Comm-LogProperResources and Comp-LogProperResources).

With the purpose of measuring the communication and computation times of each application, we have inserted a barrier (MPI_Barrier) that allow us to properly measure them. The tables of Figure 4 and Figure 5 show the overhead in percentage introduced by each message logging approach with the barriers and also without them. The executions without barriers are faster than the execution with barriers and we present both overheads in order to prove that the measures taken are consistent when removing them and executing the original versions.



Overheads With Barriers				Overheads Without Barriers			
# Process	Ovh-ShResources-NoAF	Ovh-ShResources	Ovh-ProperResources	# Process	Ovh-ShResources-NoAF	Ovh-ShResources	Ovh-ProperResources
16	39,91	38,79	38,25	16	24,24	24,41	26,25
24	55,79	50,99	50,02	24	40,31	39,60	35,52
32	56,33	37,68	25,74	32	50,28	47,66	42,94
40	75,11	65,36	51,97	40	56,07	57,63	44,82
48	89,70	79,68	65,60	48	80,68	69,55	58,41
56	101,65	82,75	61,83	56	83,45	73,34	52,41
64	87,01	74,10		64	179,03	138,30	

Fig. 5. Characterization results and Overhead Analysis of the Laplace Solver.

In Figure 4 we can observe how the computation times when using the version with own resources is lower. Even when the application becomes communication bound (56 processes) the logging version with own resources behaves better than the other versions. We do not show results of the version with own resources with 64 processes because our test environment has 64 cores, and we have to save 8 cores (1 per node) for the logger threads.

The tables of Figure 4 reflect what we have observed when characterizing the application, using message logging with own resources for the logger threads introduces less overhead in almost all cases (except with 16 cores without barriers). At best, we have reduced 25% overhead when comparing the own resources version with the version with shared resources and affinity. We can also observe that when increasing the number of processes without increasing the problem size, the overhead added becomes bigger.

Figure 5 shows the execution of the Laplace Solver. As was in the previous experiment, here we can observer how the computation times are lower when using the version with own resources.

The tables of Figure 5 reflect again what we have observed when characterizing the application, using message logging with own resources for the logger threads introduces less overhead in almost all cases (except with 16 cores without barriers).At best, we have reduced 20% overhead when comparing the own resources version with the version with shared resources and affinity.

As we have observed, in both applications the computation time of the versions with FT with own resources is lower than the versions with shared resources, but is not equal to the version without message logging. This is because

when logging is activated and a process call MPI_Irecv, this process should save the request, re-send the message to its logger thread and free the request when the message was totally received, thus there is an slight increase in computation.

6 Conclusions

The main contribution of this paper consists on analyzing possible configurations of the pessimistic receiver-based logging approach in order to find the most suitable according to application behavior. This is done by characterizing the parallel application (or a small kernel of it) obtaining the computation and communication times and the disturbance caused by the logging approaches. Our initial results have demonstrated that saving resources for the FT tasks reduces overheads and also allows us to save memory for a message log buffer. In our experimental validation we have obtained 25% overhead reduction at best.

Future work will extend the analysis made in this paper to a bigger set of applications. We will focus on obtaining traces of parallel applications and use them to find the FT configuration that will be more suitable to them. We will also analyze the relationship between message sizes and logging overheads, in order to determine the number of resources that should be save for FT tasks, because with bigger message sizes delays could increase.

References

1. Nielsen, I., Janssen, C.L.: Multicore challenges and benefits for high performance scientific computing. Sci. Program. (2008) 277–285
2. Bouteiller, A., Herault, T., Bosilca, G., Dongarra, J.J.: Correlated set coordination in fault tolerant message logging protocols. (2011) 51–64
3. Bouteiller, A., Ropars, T., Bosilca, G., Morin, C., Dongarra, J.: Reasons for a Pessimistic or Optimistic Message Logging Protocol in MPI Uncoordinated Failure Recovery. (2009) 229–236
4. Bouteiller, A., Bosilca, G., Dongarra, J.: Redesigning the message logging model for high performance. Concurr. Comput. : Pract. Exper. (2010) 2196–2211
5. Fialho, L., Rexachs, D., Luque, E.: What is missing in current checkpoint interval models? 2012 IEEE 32nd International Conference on Distributed Computing Systems (2011) 322–332
6. Lemarinier, P., Bouteiller, A., Herault, T., Krawezik, G., Cappello, F.: Improved message logging versus improved coordinated checkpointing for fault tolerant mpi. 2012 IEEE International Conference on Cluster Computing (2004) 115–124
7. Meyer, H., Rexachs, D., Luque, E.: Radic: A fault tolerant middleware with automatic management of spare nodes. The 2012 International Conference on Parallel and Distributed Processing Techniques and Applications, July 16-19, Las Vegas, USA (2012) 17–23
8. Santos, G., Fialho, L., Rexachs, D., Luque, E.: Increasing the availability provided by radic with low overhead. IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER '09. (2009) 1–8
9. Hursey, J., Squyres, J., Dontje, T.: Locality-aware parallel process mapping for multi-core hpc systems. IEEE International Conference on Cluster Computing (2011) 527–531

Optimizing Multi-Core Algorithms for Pattern Search

Veronica Gil-Costa[1,2], Cesar Ochoa[1] and Marcela Printista[1,2]

[1] LIDIC, Universidad Nacional de San Luis,
Ejercito de los Andes 950, San Luis, Argentina
[2] CONICET, Argentina

{gvcosta,mprinti}@unsl.edu.ar, elcessar@gmail.com

Abstract. The suffix array index is a data structure formed by sorting the suffixes of a string into lexicographic order. It is used for string matching, which is perhaps one of those tasks on which computers and servers spend quite a bit of time. Research in this area spans from genetics (finding DNA sequences), to keyword search in billions of web documents, to cyber-espionage. The attractiveness is that they are completely "array based" and have some benefits in terms of improving the locality of memory references. In this work we propose and evaluate the performance achieved by a static scheduling algorithm for suffix array query search. We focus on multi-core systems 32-core system. Results show that our proposed algorithm can dramatically reduce the cost of computing string matching.

Keywords: Pattern query search, multi-core, scheduling.

1 Introduction

New powerful processors and cheap storage allow to considerate alternative models for information retrieval other than the traditional one of a collection of documents indexed by keywords. One of these models is the full text model. In this model documents are represented by either their complete full text or extended abstracts. The user expresses his/her information need via words, phrases or patterns to be matched for and the information system retrieves those documents containing the user specified strings. While the cost of searching the full text is usually high, the model is powerful, requires no structure in the text, and is conceptually simple [9].

To reduce the cost of searching a full text, specialized indexing structures are used. Suffix arrays or pat arrays [9] are sophisticated indexing structures which take space close to the text size [10]. They efficiently perform phrases searching or complex queries such as regular expressions. In addition, suffix arrays can be used to index texts other than occidental natural languages, which have clearly separated words that follow some convenient statistical rules [9]. Examples of these applications include computational biology (ADN or protein strings), music retrieval (MIDI or audio files), oriental languages (Chinese, Korean, and others), and other multimedia data files.

In this work we evaluate the performance achieved by the suffix array index on a multi-core system, as multi-core systems have increasingly gained importance in high performance computers. Furthermore, we propose a scheduling algorithm devised to reduce access memory conflicts by increasing data locality and concurrency. We design our proposed scheduler to divide the pattern query search process into two steps. In the first step, all threads collaborate to classify queries into at most four groups. Each query is assigned to a particular group according to which part of the index is going to be required to process that query. In the second step, threads are assigned to process just one group of queries.

There are a number of methods proposed to schedule tasks over a set of cores. For instance, [8] shows how to reduce task pool overheads, [1] proposes a scheduler using information to determine the number of processors assigned to execute a job. The work presented in [7] allows stealing tasks from a queue using a specific data structure. The work in [15] presented a scheduling algorithm for metric space searches. The proposals are based on local or global index partitions. In the former, the database is split among threads and single index is build for each thread. Then all queries are processed by all threads. In the global index partition a single index is built and then the index is evenly distributed among threads. We did not find any related work in the literature for multi-core systems applied to problems with the features of pattern searches with suffix arrays.

The paper is organized as follows. The suffix array index and the parallel approach for multi-core systems are presented in Section 2. Our proposed scheduler algorithm is described in Section 3. Section 4 details the results obtained. Finally, conclusions are drawn in Section 5.

2. Suffix Array Data Structure

String matching is perhaps one of the most studied areas of computer science [4]. This is not surprising, given that there are multiple areas of application for string processing, being information retrieval and computational biology among the most notable nowadays. The string matching problem consists of finding some string (called the pattern) in some usually much larger string (called the text).

Many index structures have been designed to optimize text indexing [12,13]. In particular, suffix arrays [9] has already 20 years old and it is tending to be replaced by indices based on compressed suffix arrays or the Burrows-Wheeler transform [2,3], which require less memory space. However, these newer indexing structures are slower to operate. A suffix array is a data structure that is used for quickly searching for a keyword in a text database. Suffix arrays only provide efficient querying if T plus the index require less main memory than is available on the host computer, because random accesses are required to the index and the text

Essentially, the suffix array is a particular permutation on all the suffixes of a word. Given a text T[1..n] over an alphabet $\sum$, the corresponding suffix array SA[1..n] stores pointers to the initial positions of the text suffixes. The array is sorted in lexicographical order of the suffixes. As shown in Fig. 1 the text is “Performance$” and the symbol $\$ \notin \sum$ is a special text terminator, that acts as a sentinel. Given an

interval of the suffix array, notice that all the corresponding suffixes in that interval form a lexicographical subinterval of the text suffixes.

	1	2	3	4	5	6	7	8	9	10	11	12
T:	P	e	r	f	o	r	m	a	n	c	e	$

i	text suffix	SA[i]
1	$	12
2	ance$	8
3	ce$	10
4	e$	11
5	erformance$	2
6	formance$	4
7	mance$	7
8	nce$	9
9	ormance$	5
10	Performance$	1
11	rmance$	6
12	rformance$	3

Fig. 1: Suffix array for the example text "Performance$".

```
Search(SA, pattern X, pattern size L)
left =1
right=n
while (left<right)
    center = (left+right)/2
    strncpy(buff, T+SA[center-1], L)
    if (cmp=strcmp(X,buff) == 0)
       return (center-1);
    else if (cmp<0)
            right =center-1
         else
            left  = center+1
```

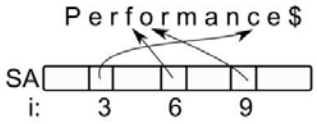


Fig.2: Search algorithm for a pattern query X of size L.

The suffix array stores information about the lexicographic order of the suffixes of the text T, but there is no information about the text itself. Therefore, to search for a pattern X[1..m], we have to access both the suffix array and the text T, with length |T|=n. Therefore, if we want to find all the text suffixes that have X as a prefix—i.e., the suffixes starting with X, and since the array is lexicographically sorted, the search for a pattern proceeds by performing two binary searches over the suffix array: one with the immediate predecessor of X, and other with the immediate successor. We obtain in this way an interval in the suffix array that contains the pattern occurrences. Fig. 2 illustrates how this search is carried out. Finally, Fig. 3 illustrates the code of the sequential search of NQ pattern queries. To this end, all patterns of length L are

loaded into main memory (to avoid the interference of disk access overheads). The algorithm scans sequentially the patterns array using the next() function, and for each pattern X the SA search algorithm is executed.

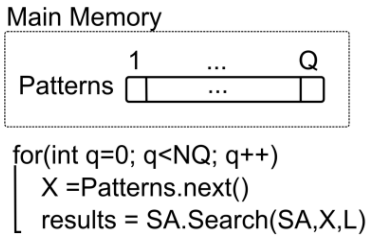


Fig.3: Sequential search algorithm

2.1 Multi-core Approach

Parallelization for shared memory parallel hardware is expected to be both, the most simple and less scalable parallel approach to a given code [13]. It is simple, due to the shared memory paradigm requires few changes in the sequential code and it is almost straightforward to understand for programmers and algorithm designers.

In this work we use the OpenMP[OMP] library, which allows for a very simple implementation of intra-node shared memory parallelism by only adding a few compiler directives. The strategy is to distribute the NQ pattern queries between execution cores. The OpenMP parallelization is made with some simple changes to the sequential code. To avoid read/write conflict, every thread should have a local *buff* variable which stores a suffix of length L, also local *left*, *right* and *cmp* variables which are used to decide the subsection of the SA where to continue the search (see Fig. 2 above). To avoid using a critical section which incurs into an overhead (delay in execution time), we replace the result variable of Fig. 3 by an array *results*[1..NQ]. Therefore, each thread stores the results for the pattern Xi into *results*[i]. The "for" instruction is divided among all threads by means of the "#pragma omp for" directive. Fig. 4 shows the threaded execution using the OpenMP terminology. Also, the sched_setaffinity function is used in this implementation to obtain performance benefits and to ensure that threads are allocated in cores belonging to the same sockets.

```
omp_set_num_threads(NT)  //Number of threads
#pragma omp parallel private(tid,X) shared(L,results)
 tid = omp_get_thread_num()
#pragma omp for
   for (int q=0; q<NQ; q++)
      X = Patterns[q]
      results[q]  = SA.Search(SA,X,L)
```

Fig. 4: Parallel search algorithm.

3. Scheduling Pattern Queries

In this section we present a static scheduling algorithm devised to improve data locality for multi-core. To this end, we divide the pattern query processing operation into two steps. In the former, all threads work to classify queries into four groups. In the last step, threads search the queries using the Suffix Array index. Both steps are separated by a barrier synchronization to avoid data corruption.

Fig. 5 shows our proposed static scheduling algorithm which aims to prune those parts of the index which will not be acceded by any query. In this figure, we show four local queues assigned to different parts of the index. The partition criterion is set according to the SA centers. Then, all threads compare the incoming queries with the centers (three centers) of the suffix array. According to the comparison result, each query is assigned to a local queue. Pattern queries matching any of the three centers are considered solved. Therefore, they are not included in any local queue for processing in the second step.

After all incoming queries are pre-processed, we assign threads to each part of the index. The basic idea is to allocate threads according to the workload. To this end, we compute the minimum number of threads required by each partition:

```
minThreads[i] = (Local_queue[i].size()/totalQ)*NT

if ( tid < minThreads[i]  )
    idGroup = i
```

where Local_queue[**i**].size() is the number of queries assigned to the index partition **i**, with **i**={1,2,3,4}. totalQ is the total number of queries that have to be processed in the second step, and NT is the number of threads. Then, each thread determines its index partition (idGroup) using its thread identifier (tid). Notice that in Fig. 5 the fourth index partition has no query assigned to the local queue. Therefore, no thread is allocated to this partition.

If the scheduling algorithm is executed with only two threads, we divide the index suffix array in two partitions (with two local queues) using the middle center as the partition criteria. With more threads we keep at most four partitions, due to increasing the number of partitions implies comparing the query with more elements of the suffix array which leads, for most of the queries, to finish their processing process before executing step 2.

In the second step, threads remove pattern queries from the local queue and search for them in the corresponding index partition. When more than one thread is allocated to the same index partition, they work in parallel by processing one query per thread.

Our hypothesis is that by dividing the pattern query search into two steps we increase data locality on the cache memories. In the first step threads compare the queries against three elements of the suffix array. Therefore, each thread will find these elements in cache avoiding accessing main memory. In the second step, threads access one part of the index, which also tends to increase the number of elements found in cache. The effect of our hypothesis has a direct effect on running time.

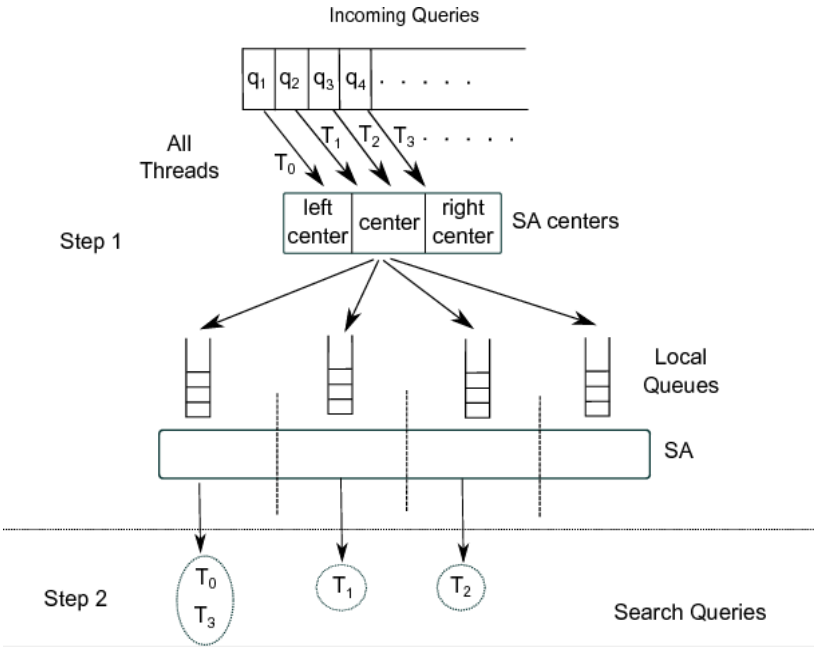


Fig. 5: Static Scheduling Algorithm.

4 Evaluation

4.1 Computing Hardware and Data Preparation

Experiments were performed on a 32-core platform with 64GB Memory (16x4GB), 1333MHz and a disk of 500GB. 2x AMD Opteron 6128, 2.0GHz, 8C, 4M L2/12M L3, 1333 Mhz Maximum Memory. Power Card, 250 volt, IRSM 2073to C13. The operating system is Linux Centos 6 supporting 64 bits. As shown in Fig. 7, we used the hwlock (Hardware Locality) tool [6] which collects all information from the operating system and builds an abstract and portable object tree: memory nodes (nodes and groups), caches, processor sockets, processor cores, hardware threads, etc. We used the OpenMP [11] library to implement the parallel codes.

To evaluate the performance of the SA index we use a 100-megabyte DNA text from the Pizza&Chili Corpus[5]. The resulting suffix array requires 801M. The text length is n=104857600. For the queries, we used 900000 random search patterns of length 10 and 30.

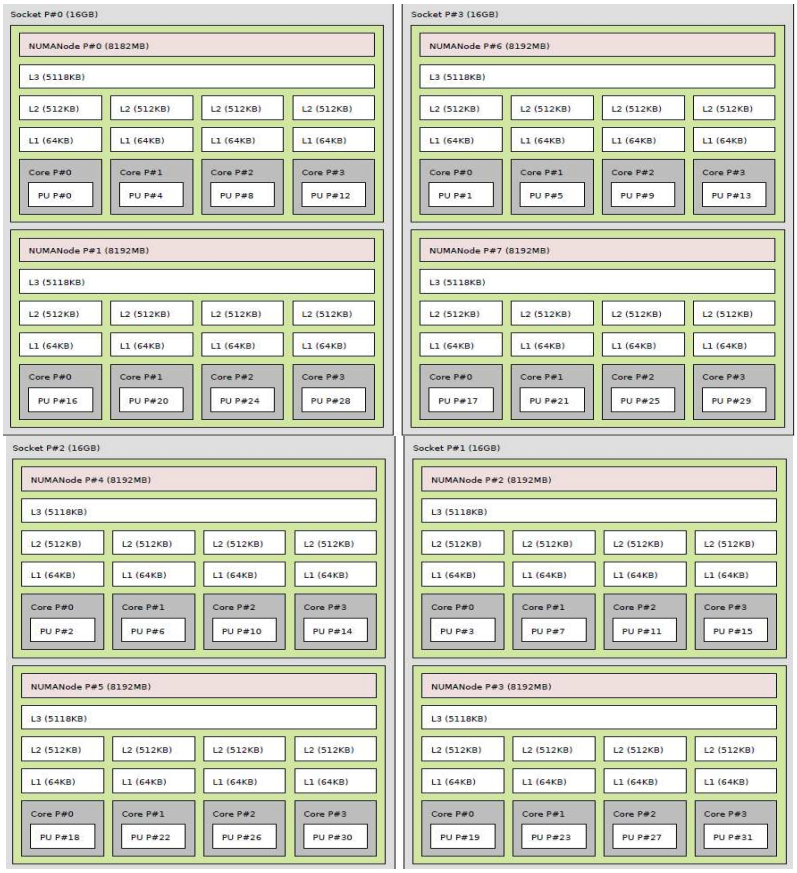


Fig. 7: Four sockets with two nodes. Each node has four cores.

4.2 Performance Evaluation

In this section we present experiment results for processing pattern queries of length L=10 and L=30 over the DNA text collection. We evaluate the baseline parallel suffix array algorithm as described in Section 2.1 and the parallel scheduler algorithm described in Section 3.

Fig. 8 and 9 show speed-up and Fig. 10 and 11 show efficiency measured as Speed-up/T, where T is the number of threads. Efficiency close to 1 indicates an optimal performance of the parallel algorithm. In all cases we use one thread per core and we execute the sched_setaffinity function.

With a small pattern length, the scheduler algorithm dramatically improves the performance of the suffix array search algorithm. The scheduler reports an improvement of 6,4% for T=2 and 77% better speed-up for T=32. Although, for T=32 the speed-up archived is 18,25 far from the optimal. These results can be

verified with efficiency reported in Fig. 10. In this last figure, the scheduler algorithm reports an efficiency close to 1 for T=2,4 and 8. Then with more threads, efficiency goes down to 60%. On the other hand, the baseline suffix array algorithm reports an efficiency close to 85% for T=2 and T=4, but it is drastically reduced to 10% for T=32.

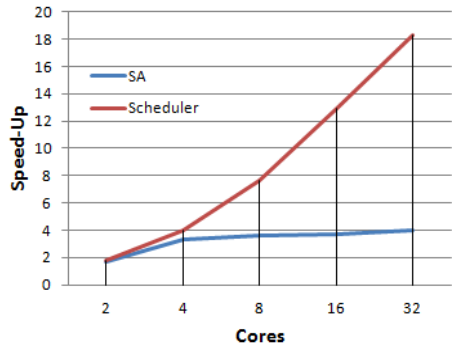


Fig. 8 Speed-Up achieved for pattern length L=10.

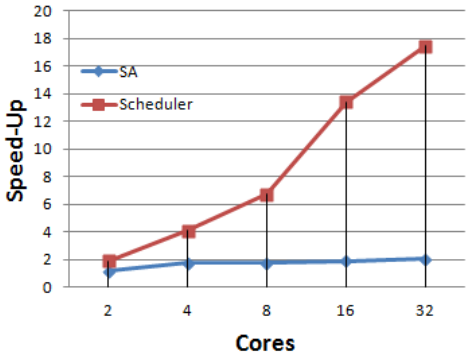


Fig. 9 Speed-Up achieved for pattern length L=30.

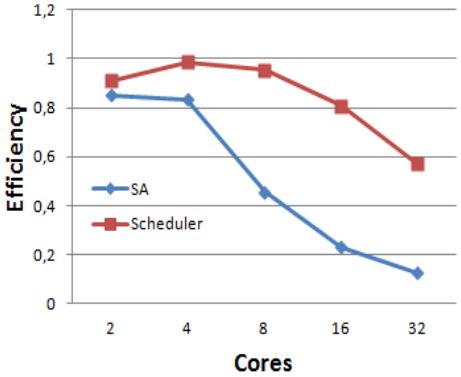


Fig.10 Efficiency for pattern length L=10.

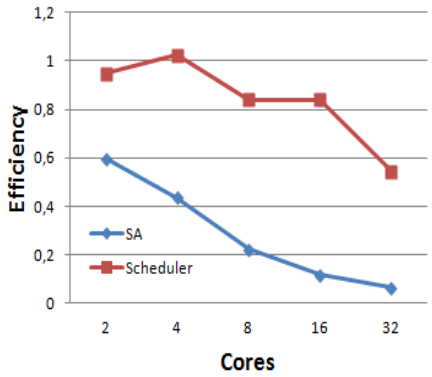


Fig. 11 Efficiency for pattern length L=30.

With a larger pattern length of L=30 (Fig. 9), the improvement reported by the scheduler algorithm over the baseline suffix array search algorithm is larger. In this case, the scheduler presents an improvement of 37% for T=2, and for T=32 the improvement is about 88%. Again, the efficiency showed in Fig. 11 confirms that the proposed scheduler algorithm reports better performance than the baseline.

For L=10, the first step consumes 15% of the total running time and the second step consumes 85% of the total running time. But with L=30, the first step consumes 12% of the total running time and the second step consumes 88% of the total running time.

In Table 1, we show the efficiency obtained by the scheduler algorithm in both processing steps. We measure the efficiency as the amount of comparisons performed by each thread (cmp) divided by the maximum number of comparisons (Max(cmp)). The formula is Σ cmp/Max(cmp)/T, where T is the number of threads. Efficiency close to one indicates that all threads perform the same amount of comparisons.

	Efficiency- Scheduler		
Threads	Step 1	Step 2	SA Baseline
2	0,99	0,751	0,99
4	0,99	0,750	0,99
8	0,99	0,748	0,99
16	0,99	0,748	0,99
32	0,99	0,748	0,99

Table 1: Efficiency reported by the Scheduling algorithm in each processing steps.

Table 1shows that the scheduler algorithm presents a balance workload in the first step, as all threads reports an efficiency of 99%. In the second step, efficiency goes down to 74% in average. Namely, some threads support a workload 30% higher.

The baseline suffix array parallel algorithm also presents an efficiency of 99%. In other words, all threads perform almost the same amount of comparisons. This is due to all queries are evenly processed by threads.

5 Conclusions

In this paper we presented a two step scheduler algorithm to improve suffix array pattern query searches. The design of the algorithm is devised to improve the data locality which has a direct effect on running time. It also prunes part of the index that will not be accessed by a given query. In the first step, the algorithm classifies and put pattern queries into at most four groups. Then, threads are assigned to process the queries of each group according to the workload.

Results show that our proposal improves the performance of the baseline suffix array index. But they also show that there is room for more research work to improve speed-up. Thus, as future work we plan to study dynamic scheduling algorithms. Namely, threads can be initially assigned to one group of queries and then during execution time, they can be moved to other groups. In this case, we must evaluate whether the cost of moving threads (cost of cache refreshing) is negligible. We also plan to study the effect of processing queries with different arrival rates.

References

1. K. Agrawal and Y. He and E. Leiserson. Adaptive work stealing with parallelism feedback. In Principles and Practice of Paralle Computing, pages 112-120. 2007.
2. Donald Adjeroh, Tim Bell, and Amar Mukherjee. The Burrows-Wheeler Transform: Data Compression, Suffx Arrays, and Pattern Matching. Springer, 2008.
3. Michael Burrows and David J. Wheeler. A block-sorting lossless data compression algorithm. Research Report 124, Digital Systems Research Center, Palo Alto California, May 1994.
4. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to Algorithms (3rd ed.). MIT Press, 2009.
5. Ferragina, P. and Navarro, G. 2005. The Pizza&Chili corpus — compressed indexes and their testbeds.
6. The Portable Hardware Locality (hwloc): http://www. open-mpi. org/ projects/ hwloc/
7. D. Hendler and N. Shavit. Non-blocking steal-half work queues. In PODC, pages 280-289. 2002.
8. R. Hoffmann and M. Korch and T. Rauber. Performance Evaluation of Task Pools Based on Hardware Synchronization. In Supercomputing Conference. 2004.
9. Manber, U. and Myers, G., “Suffix arrays: A new method for on-line string searches”, SIAM J. Computation, vol.22, no.5, pp.935-948, 1993.
10. M. Marin and G. Navarro, "Distributed Query Processing using Suffix Arrays", In International Symposium on String Processing and Information Retrieval (SPIRE 2003), Manaus, Brazil, Oct. 8-10, Lecture Notes in Computer Science 2857, pp. 311-325, Springer, 2003.
11. OpenMP Architecture Review Board, OpenMP Application Program Interface - Version 3.1, July 2011.Available athttp://openmp.org/wp/
12. Jens Stoye. Suffix tree construction in ram. In Encyclopedia of Algorithms. Springer, 2008.
13. Fernando G. Tinetti, Sergio M. Martin. Sequential Optimization and Shared and Distributed Memory Optimization in Clusters: N-BODY/Particle Simulation. Parallel and Distributed Computing and Systems -PDCS. 2012
14. Peter Weiner. Linear pattern matching algorithms. In Proceedings of the 14th Annual Symposium on Switching and Automata Theory, pages 1-11, 1973.
15. Gil-Costa Veronica, Barrientos Ricardo, Marin Mauricio and Bonacic Carolina. Scheduling Metric-Space Queries Processing on Multi-Core Processors. Euro-PDP 2010 - The 18th Euromicro International Conference on Parallel, Distributed and Network-Based Computing. IEEE. Pp. 187-194. Pisa, Italy February 17th to 19th, 2010. ISBN: 978-0-7695-3939-3

A tool for detecting transient faults in execution of parallel scientific applications on multicore clusters

Diego Montezanti[1], Enzo Rucci[1], Dolores Rexachs[2], Emilio Luque[2], Marcelo Naiouf[1] and Armando De Giusti[1,3],

[1] III LIDI, Facultad de Informática, Universidad Nacional de La Plata
Calle 50 y 120, 1900 La Plata (Buenos Aires), Argentina
{dmontezanti, erucci, mnaiouf, degiusti}@lidi.info.unlp.edu.ar
[2] Computer Architecture and Operating Systems Department, Universitat Autònoma de Barcelona
Campus UAB, Edifici Q, 08193 Bellaterra (Barcelona), Spain
{dolores.rexachs, emilio.luque}@uab.es
[3] Consejo Nacional de Investigaciones Científicas y Tecnológicas

Abstract. Transient faults are becoming a critical concern among current trends of design of general-purpose multiprocessors. Because of their capability to corrupt programs outputs, their impact gains importance when considering long duration, parallel scientific applications, due to the high cost of re-launching execution from the beginning in case of incorrect results. This paper introduces SMCV tool which improves reliability for high-performance systems. SMCV replicates application processes and validates the contents of the messages to be sent, preventing the propagation of errors to other processes and restricting detection latency and notification. To assess its utility, the overhead of SMCV tool is evaluated with three computationally-intensive, representative parallel scientific applications. The obtained results demonstrate the efficiency of SMCV tool to detect transient faults occurrences.

Keywords: Transient fault, parallel scientific application, soft error detection tool, message content validation.

1 Introduction

The increase in the integration scale, in order to improve computing performance of current processors, as well as the growing size of the computer systems (towards upcoming exascale), are factors that make reliability an important issue. Particularly, transient faults, also named soft errors, are becoming a critical concern because of their capability to affect program correctness [1].

A transient fault is caused by interference from the environment, such as electromagnetic radiation, overheating or input power variations. It can alter signal transfers, register values, or some other processor component, temporarily inverting one or several bits of the affected hardware element [2]. Although short-lived transient faults do not cause permanent physical damage to the processor, depending on the moment or specific location of the occurrence, they may corrupt computations, resulting in either control flow faults or data faults that may propagate and cause

incorrect program execution [3][4]. Soft errors have led to costly failures in high-end systems in recent years [5][6].

The increasing number of transistors per chip involves lower voltage thresholds and higher internal operating temperatures. As a consequence, the vulnerability of the entire chip to transient faults (i.e. the soft error rate) is expected to increase significantly [7][8]. As soft errors can cause serious reliability problems, all general purpose microprocessors (especially those that form part of high availability systems) should employ fault-tolerance techniques to ensure right operation.

The impact of transient faults becomes more significant in the context of High Performance Computing (HPC). Since the year 2000, error reports due to transient faults in large computers or server groups have become more frequent [5][6]. Moreover the impact of the faults becomes more relevant in the case of long-duration applications, given the high cost of re-launching execution from the beginning. These factors justify the need for a set of strategies to improve the reliability of high-performance computation systems.

Historically, transient faults have been a design concern in critical environments, such as flight systems or high-availability servers. To face them, additional hardware is introduced, varying from watchdog co-processors to redundant hardware threads [9][10][11][12][13][14]. Storage devices, memories, caches have efficient built-in mechanisms such as Error Correcting Codes (ECC´s) or parity bits, capable of detecting or even correcting this type of faults [4]. In practice, these techniques are costly or impossible to apply to processor elements [3] and they result inefficient in general purpose computers, due mainly to the high cost of designing, developing and verifying redundant custom hardware [1]. In this context, the faults that affect processor registers are a concern. In addition, as architectural trends point toward multicore designs, there is substantial interest in adapting such parallel hardware resources for transient fault tolerance.

To provide protection with lower (or zero) hardware costs, software-only approaches have been proposed [3][4]. Despite having some limitations (they have to execute additional instructions and are unable to examine microarchitectural state), software-only techniques have shown promise, in the sense that they can significantly improve reliability with reasonable performance overhead [15][16][17]. This characteristic makes software-redundancy-based strategies to be the most appropriate for general purpose computational systems.

Most software-duplication based techniques are designed for serial programs. From this standpoint, a parallel application can be viewed as a set of sequential processes that have to be protected from the consequences of transient faults adopting the software-based techniques.

MPI [18] is currently the de facto standard that defines an API for a message-passing parallel programming model. MPI is designed for achieving portable high-performance communication in parallel applications. However, while the current parallel computing systems are improving their robustness, the MPI specification does not fully exploit the capabilities of the current architectures [19][20].

Because the addition of reliability features in communication increases processing and resource overheads, MPI offers limited fault-handling facilities. Despite the fact that MPI processes may fail because of any external fault (e.g. processor, network or power failures), detection of such faults is not defined in the standard.

According to such scenario, in recent past SMCV methodology has been proposed [21], which is a software-only approach specifically designed for the detection of transient faults in message-passing parallel scientific applications that execute on multicore clusters.

In order to facilitate the usability of SMCV methodology, this paper presents SMCV tool, which is a library of modified MPI functions and data types with extended functionality for fault detection by comparison upon sending, message contents duplication upon reception, and concurrency control between replicas. SMCV tool has the goal of helping programmers and users of parallel scientific applications to achieve reliability in their executions, obtaining correct final results or, at less, reporting the silent fault occurrence and avoiding its consequences by leading to a safe-stop state. This avoids the unnecessary and costly wait until execution finishes, allowing application re-launching after a restricted delay due to latency of detection. This is an important feature, owing to the long duration executions of such applications.

To estimate the impact of SMCV tool on performance of message-passing parallel scientific applications, and in order to evaluate the convenience of its utilization, a set of experiments was made, using three benchmark parallel applications: matrix multiplication [22]; solution to Laplace's equation [23]; and DNA sequence alignment [24]. With these experiments, the performance of the tool was evaluated for various problem sizes using different number of processes, obtaining 93.7 maximum and 24.3 average percent overhead in absence of faults. As explained further on, at least two executions of the original application and final results comparison are needed to determine if a transient fault has occurred when no fault tolerance strategy is employed by the system. Accordingly, these results demonstrate the efficiency of SMCV tool.

The rest of this paper is organized as follows: Section 2 discusses related works. Section 3 reviews the theoretical context of transient faults. Section 4 and Section 5 describes SMCV methodology and SMCV tool respectively. In Section 6, the experimental work carried out is described, whereas Section 7 presents and analyzes the obtained results. Finally, Section 8 presents the conclusions and future lines of work in relation to this paper.

2 Related Works

Redundancy techniques can be broadly classified into two kinds: hardware-based and software-based. There have been various implementations of software-only, hardware-only, and hybrid techniques for transient fault mitigation [3][4].

All hardware-based approaches require the addition of some new hardware logic to meet redundancy requirements. Several researchers have also made use of multiplicity of hardware blocks readily available on multithreaded/multicore architectures to implement computation redundancy [10][11][12][14].

Fault tolerance based on software replication is a well-populated field with decades of history. Their main advantage is that they do not require any additional hardware. Among the purely software solutions, PLR [1] is a process replication-based one.

Other software-only techniques for transient fault detection are the compiler-based ones. At compile time, they insert redundant computations [16], control flow assertions [15] or both [4].

As regards to hybrid strategies, in [3], the authors propose a fault-tolerant typed assembly language, in an attempt to exploit the benefits of both hardware and software-based systems for fault tolerance.

All the previously mentioned proposals are designed for sequential applications. SMCV is specific for message-passing parallel scientific applications.

There are some approaches that extend MPI to implement process replicas on MPI applications for hard faults. MPI/FT [20] is an MPI-based middleware that provides additional services for detection and recovery of failed MPI processes. FT-MPI [19] specifies the semantics of a fault tolerant version of MPI and implements that specification. Whereas the two mentioned strategies provide support for failures that make a process to terminate, SMCV provides a mechanism for detecting transient faults in MPI applications improving at the same time system availability. No proposals for transient fault detection in parallel scientific applications based on message validation were found while researching for this work.

3 Background on soft errors

As aforementioned, transient faults affect system hardware elements, but their effects are observed on the program execution (assuming deterministic programs). According to these effects, they can be classified into the following categories:

- Latent Error (LE): also called benign fault, is a fault that corrupts data that are not used by the application so, despite the fault effectively happening, it does not propagate to affect the correctness of the execution and has no impact on the results.
- Detected Unrecoverable Error (DUE): is a detected error that has no possibility of recovery. DUEs are a consequence of faults that cause abnormal conditions that are detectable on some intermediate software layer level (e.g. Operating System, communication library). Normally, they cause the abrupt stop of the application.
- Time-Out Error (TO): due to fault, the program does not terminate within a given amount of time.
- Silent Data Corruption (SDC): is the alteration of data during the execution of a program that does not cause any abnormal condition and goes undetected by system software. Its effects are silently propagated to corrupt final results. This is the worst case, because the application appears to execute correctly buy silently produce incorrect output [1].

4 SMCV Methodology Overview

SMCV is a detection strategy based on validating the contents of the messages to be sent in deterministic parallel scientific applications. In particular, SMCV intercepts

faults that produce TOs and SDCs. Under this approach, each application process is duplicated and the process and its replica run concurrently, which requires a synchronization mechanism. When a communication is to be performed (point-to-point or collective), the process temporarily stops execution and waits for its replica to reach the same point. Once there, all fields from the message to be sent are compared to validate that the contents of both threads are the same. Only if this proves true, one of the threads sends the message, ensuring that no corrupt data are propagated to other process. The recipient(s) of the messages stop upon reception and remain on hold. Once received, it copies the contents of the message to its replica and both processes continue with their computation. Finally, when application execution finishes, the obtained results are checked to detect faults that may have occurred after communications ended, (i.e. the serial part of the application).

Figure 1 shows SMCV detection outline whereas Figure 2 shows the SMCV behavior in presence of transient faults. More details about SMCV methodology can be found in [21].

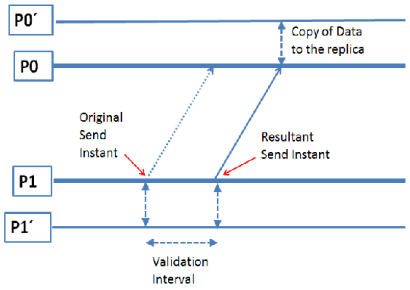


Fig. 1. SMCV detection outline.

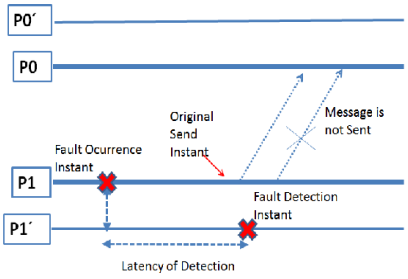


Fig. 2. SMCV behavior in presence of transient faults.

5 SMCV Tool

5.1 Description

To implement SMCV methodology, SMCV tool was developed. It consists of a library of modified MPI functions and data types that can be used in MPI applications developed using C language. SMCV library redefines MPI functions and data types with only on syntactic change (the `MPI` prefix is replaced with `SMCV`). In turn, it adds two new functions: `SMCV_Call` and `SMCV_Validate`. For threads replication and synchronization, Pthreads functions are used. MPI functions redefinition is necessary to provide transient fault detection in a transparent way to applications code and their programmers. This implies application source code modification and recompiling.

5.2 Basic Functions

MPI standard defines six basic functions [25]. The SMCV library core consists of the six redefined MPI basic functions and two other. These eight functions are enough to develop a wide range of parallel applications that are able to detect transient faults. SMCV basic functions are described below:

SMCV_Init. Initiate a SMCV environment.

SMCV_Finalize. Terminate a SMCV environment.

SMCV_Comm_size. Determine number of processes.

SMCV_Comm_rank. Determine process identifier.

SMCV_Call. Create a new thread that executes the code to be validated.

SMCV_Send. Synchronize the process and its replica. The second to reach the synchronization point compares all the fields of the message to be sent (byte to byte). If all fields match, the first thread sends the message. Once sent, both threads continue with their execution. If any field differs, a safe-stop is produced because a SDC has occurred. Moreover, there is a (configurable) time for the second thread to reach the synchronization point, in order to be able to intercept TOs.

SMCV_Recv. Synchronize the process and its replica. The first to reach the synchronization point receives the message and remains on hold. When the second thread arrives, it copies the contents of the message received. After that, both threads continue with their execution. Like `SMCV_Send`, there is a (configurable) time for the second thread to reach the synchronization point, in order to be able to intercept TOs.

SMCV_Validate. Synchronize the process and its replica. The second to reach the synchronization point compares both threads' final result (byte to byte). If the final results match, the threads continue with their execution. Otherwise, a safe-stop is produced because a SDC has occurred. Like `SMCV_Send`, there is a (configurable) time for the second thread to reach the synchronization point, in order to be able to intercept TOs.

5.3 Usage

The next steps must be followed to incorporate SMCV features in MPI application code:

1. Replace MPI header with SMCV header.
2. Encapsulate the code to be validated (data and instructions) in a `void *` function.
3. Make a call to `SMCV_Call` function passing the previously defined function to it as an argument.
4. Replace `MPI` prefix with `SMCV` in all MPI functions and data types.
5. Make a call to `SMCV_Validate` in order to validate the application final result.

Figure 3 shows an example of how to adapt and MPI application in order to incorporate SMCV features.

```c
#include <mpi.h>
int main (int argc, char **argv)
{
   MPI_Init();
   /* Process data, instructions
and MPI functions */
   MPI_Finalize();
   return 0;
}
```

```c
#include <smcv.h>
int main (int argc, char **argv)
{
   SMCV_Init();
   SMCV_Call(&smcv_process)
   SMCV_Finalize();
   return 0;
}
void * smcv_process () {
   /* Thread data, instructions
and SMCV functions */
   SMCV_Validate();
}
```

Fig. 3. Example of how to adapt a MPI application in order to incorporate SMCV features. Left: MPI application source code. Right: SMCV-adapted MPI application source code.

6 Experimental Work

6.1 Architecture Used

Experimental work was carried out on a cluster of Blade multicores with four blades. Each blade has two quad core Intel Xeon e5405 2.0GHz processors with 6Mb L2 cache (shared between pairs of cores) and 10 Gb RAM memory (shared between both processors). The operating system is GNU/Linux Debian 6.0.7 (64 bits, kernel version 2.6.32) and the MPI library used is OpenMPI (version 1.6.4).

6.2 Benchmark Applications Used

Three benchmark parallel applications were selected: matrix multiplication [22]; solution to Laplace's equation [23]; and DNA sequence alignment [24]. These benchmark applications were selected because of three main reasons: first, they are well-known, representative scientific applications; second, they are computationally intensive; and third, they have three different communication patterns: Master-Worker, Single-Program-Multiple-Data (SPMD) and Pipeline, respectively.

Tests were carried out using MPI and SMCV versions of the three selected benchmark applications. The steps described in Subsection 5.3 were followed to incorporate SMCV features to original applications' codes. Finally, because SMCV was especially designed to be used in context of HPC applications, the `-O` optimization level was used at compile time.

6.3 Tests Carried Out

Benchmark applications were tested using different number of processes: *P={4, 8, 16}*. Various problem sizes were used for each application: *N={2048, 4096, 8192,*

16384} for matrix multiplication; *N={4096, 8192, 16384}* for solution to Laplace's equation and *N={65536, 131072, 262144, 524288}* for DNA sequence alignment. At most four processes were mapped by node, which means that in original applications execution only four cores of each node were used. In the case of SMCV applications, all the cores of each node were used (the replicas execute on available cores). Each experiment was run five times and the results were averaged to improve stability.

7 Results

To assess the incidence of SMCV tool over the applications performance when escalating the problem and/or the architecture, the *Overhead* metric is analyzed. The overhead is a consequence of the processes duplication, the synchronization with the replicas, the comparison and duplication of the messages contents and the final validation of the results. In addition, the processes duplication increases contention for system resources. Equation 1 indicates how to calculate this metric, where *APP_ET* is the original application execution time and *SMCV_APP_ET* is the SMCV-adapted application execution time.

$$Overhead = \frac{(SMCV\_APP\_ET - APP\_ET)}{APP\_ET} \times 100 \,. \quad (1)$$

Figures 4, 5 and 6 shows the overheads obtained with SMCV applications (matrix multiplication, solution to Laplace's equation and DNA sequence alignment, respectively) for various problem sizes using different number of processes.

The charts show that the three benchmark applications present similar behaviors. As it can be observed, overhead decreases as the problem size grows. This is due to, with larger problem sizes, applications spend more time computing than communicating and, consequently, the time required to synchronize threads and to duplicate and validate message contents reduces (in the case of matrix multiplication, data duplication produces disk-swapping when *N=16384* and *P={8,16}* and, as a consequence, overhead reduction does not remain). On the other hand, the number of messages to be sent increases as the number of processes grows. This leads to an overhead increase because time required to synchronize threads and to compare and duplicate message contents enlarges.

As mentioned above, overhead behaviors are similar, but the same does not occur with overhead values. Matrix multiplication is the application with largest overhead values. This is due to the sizes of the messages that processes send (matrix sizes go from 16MB to 1GB according to *N*), aggravated by the fact that they use collective communication operations for it. Unlike OpenMPI, SMCV library does not optimize this kind of communication operations [26]. Last, the final result of this application is a matrix and the time required to validate it is not insignificant.

Overhead values of the solution to Laplace's equation are lower than the corresponding ones to matrix multiplication. Even though processes repeatedly interchange messages (which increases the number of synchronizations), the time required to validate them reduces because of the smaller message size (they go from 16KB to 64KB depending on *N*). Another influence factor is that the final result is a single number and, in consequence, the time necessary to validate it is negligible.

DNA sequence alignment presents overhead values even lower than the corresponding ones to the solution to Laplace's equation. All the processes receive and send messages repeatedly (except the first and the last of the pipeline). Because of these messages are of fixed size and very small (136B), the time required to validate them is not significant. Like the previous case, final result validation does not demand considerable time.

In this set of experiments, SMCV tool provides fault detection with 93.7 maximum and 24.3 average percent overhead. This represents an advantage with respect to the original execution, which has to be repeated (and final results have to be compared) to ensure a correct output if a SDC does not occur. Moreover, if a SDC occurs, a third re-execution (and a new comparison) is required to pick the outputs of the runs that form a majority as the correct ones.

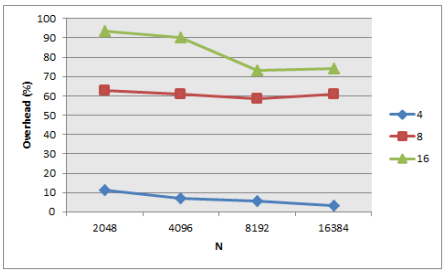


Fig. 4. Overheads obtained for SMCV-matrix multiplication for various problem sizes using different number of processes.

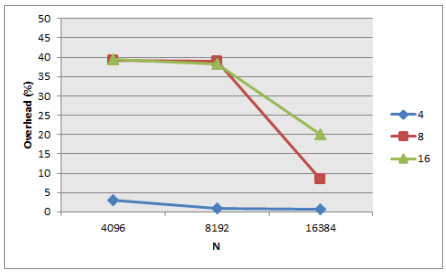


Fig. 5. Overheads obtained for SMCV-solution to Laplace's equation for various problem sizes using different number of processes.

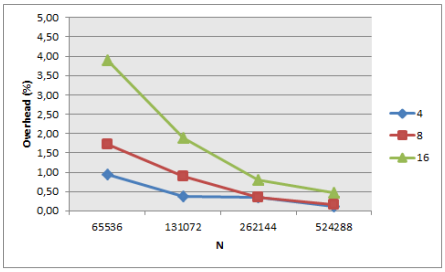


Fig. 6. Overheads obtained for SMCV-DNA sequence alignment for various problem sizes using different number of processes.

8 Conclusions and Future Work

Transient faults are becoming more frequent in large computers and their impact is higher in the case of long duration applications. In this paper, SMCV tool is presented to help programmers and users of scientific parallel applications to achieve reliability in their executions, obtaining correct final results or, at less, reporting the silent fault

occurrence within a limited time lapse and leading to a safe-stop state. Experimental results show that, when running three different benchmark parallel applications on a multicore cluster for various problem sizes and using different number of processes, SMCV tool provides fault detection with 93.7 maximum and 24.3 average percent overhead. These results demonstrate the tool's efficiency to provide transient fault detection in message-passing parallel scientific applications.

Future lines of work focus on four aspects:

- Extending current SMCV library implementation to give full support to MPI applications (at the moment it only supports blocking communication functions and some collective communication routines).
- Optimizing collective communications implementation to take benefit of MPI features, in order to minimize overheads.
- Automating the procedure to adapt the original application source code to use SMCV tool.
- Emulating non-deterministic functions, to extend SMCV methodology for giving support to transient fault detection in non-deterministic MPI scientific applications.

References

1. Shye, A., Blomstedt, J., Moseley, T., Reddi, V. J., Connors, D. A.: PLR: A software approach to transient fault tolerance for multicore architectures; IEEE Transactions on Dependable and Secure Computing. 6(2), pp. 135--148 (2009)
2. Wang, N. J., Quek, J., Rafacz, T. M., Patel, S. J.: Characterizing the Effects of Transient Faults on a High-Performance Processor Pipeline. In: Proceedings of the International Conference on Dependable Systems and Networks, pp. 61--70. IEEE Press, Florence (2004)
3. Perry, F., Mackey, L., Reis G. A., Ligatti, J., August, D. I., Walker, D.: Fault-tolerant typed assembly language. In: Proceedings of the 2007 ACM SIGPLAN conference on Programming language design and implementation, pp. 42--53. ACM Press, San Diego (2007)
4. Reis, G. A., Chang, J., Vachharajani, N., Rangan, R., August, D. I.: SWIFT: Software Implemented Fault Tolerance. In: Proceedings of the International Symposium on Code generation and optimization, pp. 243–254. IEEE Press, Washington DC (2005)
5. Baumann, R. C.: Soft errors in commercial semiconductor technology: Overview and scaling trends. In: IEEE 2002 Reliability Physics Tutorial Notes, Reliability Fundamentals, pp. 121 01.1--121 01.14.
6. Michalak, S. E., Harris, K. W., Hengartner, N. W., Takala, B. E., Wender , S. A.: Predicting the number of fatal soft errors in Los Alamos National Labratory's ASC Q computer; IEEE Transactions on Device and Materials Reliability. 5(3), pp. 329--335 (2005)
7. Gramacho, J., Rexachs del Rosario, D., Luque, E.: A Methodology to Calculate a Program ´s Robustness against Transient Faults. In: Proceedings of the International 2011 Conference on Parallel and Distributed Processing Techniques and Applications, pp. 645--651. WorldComp Press, Las Vegas (2011)
8. Mukherjee, S.; Weaver, C.; Emer, J.; Reinhardt, S., Austin, T.: A systematic methodology to compute the architectural vulnerability factors for a high-performance microprocessor.

In: Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 29--40. IEEE Press, San Diego (2003)
9. Mahmood, A., McCluskey, E. J.: Concurrent error detection using watchdog processors-a survey. IEEE Transactions on Computers. 37(2), pp. 160--174 (1988)
10. Reinhardt, S. K., Mukherjee S. S.: Transient Fault Detection via Simultaneous Multithreading. In: Proceedings of the 27th annual International Symposium on Computer Architecture, pp. 25--36. IEEE Press, Vancouver (2000)
11. Kontz M., Reinhardt S. K., Mukherjee S. S.: Detailed Design and Evaluation of Redundant Multithreading Alternatives. In: Proceedings of the 29th Annual International Symposium on Computer Architecture, pp. 99--110. IEEE Press, Anchorage (2002)
12. Vijaykumar T. N., Pomeranz, I. Cheng, K.: Transient-Fault Recovery using Simultaneous Multithreading. In: Proceedings of the 29th Annual International Symposium on Computer Architecture, pp. 87--98. IEEE Press, Anchorage (2002)
13. Gomaa M., Scarbrough C., Vijaykumar T. N., Pomeranz, I.: Transient-Fault Recovery for chip Multiprocessors. In: Proceedings of the 30th Annual International Symposium on Computer Architecture, pp. 98--109. IEEE Press, San Diego (2003)
14. Rotenberg E.: AR-SMT: A Microarchitectural Approach to Fault Tolerance in Microprocessors. In: Proceedings of the 29th Annual International Symposium on Fault-Tolerant Computing, pp. 84--91. IEEE Press, Wisconsin (1999)
15. Oh, N., Shirvani, P. P., McCluskey, E. J.: Control-flow checking by software signatures. IEEE Transactions on Reliability, 51(1), pp. 111--122 (2002)
16. Oh, N., Shirvani, P. P., McCluskey, E. J.: Error detection by duplicated instructions in super-scalar processors; IEEE Transactions on Reliability. 51(1), pp. 63--75 (2002)
17. Reis, G. A., Chang, J., August, D. I.: Automatic instruction level software-only recovery methods; IEEE Micro Top Picks. 27 (1), pp. 36--47 (2007)
18. Message Passing Interface Forum, http://www.mpi-forum.org/
19. Fagg, G.E., Gabriel, E., Chen, Z., Angskun, T., Bosilca, G., Pjesivac-Grbovic, J., Dongarra, J.J.: Process Fault-Tolerance: Semantics, Design and Applications for High Performance Computing; International Journal of High Performance Applications. 19(4), pp. 465--478 (2005)
20. Batchu, R., Dandass, Y., Skjellum, A., Beddhu, M.: MPI/FT: A Model-Based Approach to Low-Overhead Fault Tolerant Message-Passing Middleware; Cluster Computing. 7 (4), pp. 303--315 (2004)
21. Montezanti, D., Frati, F.E., Rexachs, D., Luque, E., Naiouf, M.R., De Giusti, A.: SMCV: a Methodology for Detecting Transient Faults in Multicore Clusters.; CLEI Electron. J. 15(3), pp. 1--11 (2012)
22. Leibovich, F., Gallo, S., De Giusti, A., De Giusti, L., Chichizola, F., Naiouf, M.: Comparación de paradigmas de programación paralela en cluster de multicores: pasaje de mensajes e híbrido. In: Anales del XVII Congreso Argentino de Ciencias de la Computación. pp. 241--250. Editorial RedUNCI, La Plata (2011)
23. Andrews, G.: Foundations of Multithreaded, Parallel, and Distributed Programming. Addison Wesley Longman, EEUU (2000).
24. Rucci, E., Chichizola, F., Naiouf, M., De Giusti, A.: Parallel Pipelines for DNA Sequence Alignment on Cluster of Multicores. A comparison of communication models.; Journal of Communication and Computer. 9(12), pp. 516--522 (2012)
25. Dongarra, J., Foster, I., Fox, G., Gropp, W., Kennedy, K., Torczon, L., White, A.: The Sourcebook of Parallel Computing. Morgan Kauffman, EE.UU. (2003)
26. Graham, R., Shipman, G.: MPI Support for Multi-core Architectures: Optimized Shared Memory Collectives. In: Proceedings of the 15th European PVM/MPIUsers' Group Meeting on Recent Advances in Parallel Virtual Machine and Message Passing Interface. pp. 130--140. Springer-Verlag Berlin (2008)

Lessons learned from contrasting a BLAS kernel implementations

Andrés More[1 2]

[1] Intel Software Argentina (Argentina Software Design Center)
andres.more@intel.com
[2] Instituto Aeronáutico Córdoba
amore@iua.edu.ar

Abstract. This work reviews the experience of implementing different versions of the SSPR rank-one update operation of the BLAS library. The main objective was to contrast CPU versus GPU implementation effort and complexity of an optimized BLAS routine, not considering performance. This work contributes with a sample procedure to compare BLAS kernel implementations, how to start using GPU libraries and offloading, how to analyze their performance and the issues faced and how they were solved.

Keywords: BLAS libraries, SSPR kernel, CPU architecture, GPU architecture, Performance analysis, Performance measurement, Software Optimization.
XIII Workshop de Procesamiento Distribuido y Paralelo

1 Introduction

With the growth of the application of general-purpose GPU techniques to compute intensive problems [1], there will be lots of domain experts trying to implement an specific math kernel operation both in CPU and GPU, applying optimizations and finally doing a performance comparison to double check if gains are relevant due the invested effort [2].

It is non trivial to identify the outcome. While CPUs are designed for general purpose, GPUs are designed for specific purposes; specific problems are well suited to GPU architecture and can be solved faster than using CPUs, but they usually need to be represented in a way parallelism is explicitly exposed. Picking the right approach will contribute to faster, more detailed problem solving in domain specific problems. It is expected that the procedure of reviewing an specific kernel implementations will be done on each release of new CPU and GPU architecture by people doing problem solving simulations. From that point of view this work contributes with a similar analysis experience. Related work is included as part of the results as a validation proof.

The rest of the document is structured as follows: this section is completed with information on the operation and system being used for the comparison, plus a quick discussion on the GPU programming model. Section 2 include details on which algorithms were executed after reviewing well-known implementations. Section 3 provide details on how the performance was contrasted, including results and related work validating our results.

1.1 Single-precision Symmetric Packed Rank-one update

The Single-precision Symmetric Packed Rank-one update (SSPR) operation computes a rank-1 update of a symmetric matrix of single precision floating point numbers. SSPR performs the symmetric rank 1 operation show in Equation 1, where α is a real scalar, x is an n element vector and A is an n by n symmetric matrix, supplied in packed form.

$$A := \alpha \times x \times x^T + A \tag{1}$$

A graphical representation of the operation is shown in Figure 1.

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ . & & . \\ . & & . \\ . & & . \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \leftarrow \begin{bmatrix} a_{11} & \dots & a_{1n} \\ . & & . \\ . & & . \\ . & & . \\ a_{n1} & \dots & a_{nn} \end{bmatrix} + \alpha \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$$

Fig. 1. Graphical representation of rank 1-update

The SSPR operation belongs to level 2 BLAS (Basic Linear Algebra Subprograms) [3] [4], as it runs over a matrix-vector combination. Matrix-vector multiplication has $O(N^2)$ complexity on the size of the matrix [5]. However, the matrix is required to be represented as a vector which contains only half of the symmetric matrix triangular, packed sequentially per column. Represented by Equation 2. In particular, the vector has a size of $(n \times (n+1)/2)$. This approach avoids redundancy and saves device memory for bigger input.

$$AP(i + (j \times (j-1)/2) = A_{ij} (\forall j \geq i) \tag{2}$$

Operation signature in the Fortran language is defined in Listing 1 and clarified in Table 1 below. Towards generality of the function, UPLO specifies if the packed triangular matrix is the upper or the lower part of the original data. INCX is the required increment to reference the vector elements in the provided vector reference. This is useful to iterate over a vector which is part of a matrix, avoiding extra buffer copies.

Listing 1. SSPR Fortran Signature

```
1  SUBROUTINE SSPR(UPLO, N, ALPHA, X, INCX, AP)
2              .. Scalar Arguments ..
3              REAL ALPHA
4              INTEGER INCX,N
5              CHARACTER UPLO
6              ..
7              .. Array Arguments ..
8              REAL AP(*),X(*)
```

Table 1. SSPR Arguments Description

Argument	Description
UPLO	Specifies whereas upper/lower triangular part of A is supplied in array
N	Specifies the order of the matrix A
ALPHA	Specifies the scalar alpha
X	Array of dimension at least $(1 + (n-1) * abs(INCX))$
INCX	Specifies the increment for the elements of X
AP	Array of dimension at least $((n * (n+1))/2)$

1.2 Testbed

The system used for experiments was an *HP Mobile Workstation EliteBook 8530w*, having integrated an *Intel CPU model T9600* [3] plus a *Quadro FX 770m GPU* [4]. Although the system is a little bit outdated and it is not state-of-the-art, the computing devices were integrated at the same time-frame and hence provide a valid point of comparison.

The relevant information from their hardware specifications are shown below. The CPU information was taken from `/proc/cpuinfo` and a custom program to access GPU attributes not yet exposed thru standard kernel interfaces.

```
cpu family: 6
model: 23
model name: Intel(R) Core(TM) 2 Duo CPU T9600 @ 2.80GHz
stepping: 10
cpu MHz: 2793
cache size: 6144 KB
fpu: yes
cpuid level: 13
flags: fpu vme de pse tsc msr pae mce cx8 apic sep mtrr
pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2
ss ht tm pbe pni dtes64 monitor ds_cpl vmx smx est tm2
ssse3 cx16 xtpr pdcm sse4_1 xsave osxsave lahf_lm
TLB size: 0 4K pages
```

[3] T9600 CPU specifications
[4] Quadro FX 770m GPU specifications

The GPU card has built-in 512MB memory, meaning that in average during executions there is about 462 MB of global memory available for computation.

```
capabilities.name = Quadro FX 770M
capabilities.totalGlobalMem = 512.00 MB
capabilities.sharedMemPerBlock = 16.00 KB
capabilities.regsPerBlock = 8192
capabilities.warpSize = 32
capabilities.memPitch = 2097152.00 KB
capabilities.maxThreadsPerBlock = 512
capabilities.maxThreadsDim = 512 512 64
capabilities.maxGridSize = 65535 65535 1
capabilities.totalConstMem = 64.00 KB
capabilities.major = 1
capabilities.minor = 1
capabilities.clockRate = 1220.70 MHz
capabilities.textureAlignment = 256
capabilities.deviceOverlap = 1
capabilities.multiProcessorCount = 4
cudaMemGetInfo.free = 462 MB
```

As depicted in Figure 2, the different in single-precision (real) floating point operations is significant. It might be expected results that choose GPUs as the winner, which will be a different assumption if the operation was using double precision floating point operations.

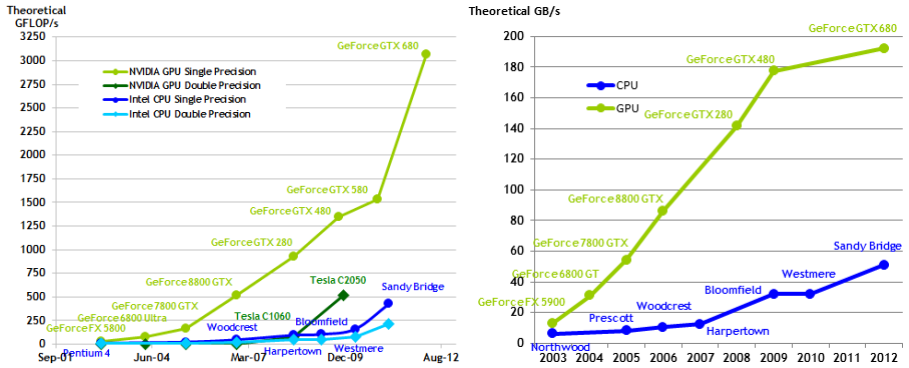


Fig. 2. FLOPs and Bandwidth Comparison between CPUs and GPUs [6]

1.3 CUDA Programming

The Compute Unified Device Architecture (CUDA) is a high performance computing architecture and programming model created by NVIDIA. The CUDA programming model gives access to GPUs instruction set to be used for general purpose computing, providing both low level and high level interfaces to simplify application.

The programming model assumes that the host system will execute kernels by asynchronous offloading to a GPU device. Memory allocation and transfer are also controlled by the developer. The programming model relies on a hierarchy of thread groups that can access per-group shared memory and can synchronize between them. An extension to the C language allows the inclusion of compute kernels that run in the GPU device. Hardware threads are part of an N-dimensional block with `N=1,2,3`.

All threads on a block are expected to share (faster than global) memory resources and to synchronize with each other. Blocks per-se are also organized in sets called grids, similar to blocks. Thread blocks perform independent computation. Each block memory hierarchy consists of local, global, shared, constant and texture memory; the latter having special addressing to better fit graphical rendering. All but shared memory is persistent across kernel executions.

GPU hardware and software capabilities are evolving fast, not only instructions per cycle have increased. Support for native arithmetic instructions have expanded the set of atomic operations and the provided math functions. The usual

`printf` capability has been recently made available, the runtime uses a special device buffer that it is transferred together with the device memory. It is worth noting that IEEE 754-2008 binary floating-point arithmetic has deviations from standard by default. Enabling strict support imposes a performance penalty.

2 Algorithms

As part of the work 4 different versions of SSPR functions were exercised, using well-known implementations as a reference. The following subsections contains details on how the versions were implemented that can be reused for any other BLAS kernel. In order to streamline analysis only support for upper matrices with 1 increments were incorporated.

2.1 CPU Sequential

Using both the original BLAS implementation [5] and the GNU Scientific Library version [6] an initial CPU sequential version was implemented as shown in Listing 2. It is worth to note that any BLAS kernel can be reviewed also on this two implementations. A naive implementation of the mathematical definition was not used as proper speedup computation requires best known sequential algorithm [7], shown in Listing 3.

Listing 2. Naive SSPR CPU Implementation

```
k=0;
for (j=0;j<n;j++)
        for(i=0;i<=j;i++) {
                ap[k] += alpha*x[i]*x[j];
                k++;
}
```

Listing 3. Optimized SSPR CPU Implementation

```
for (i = 0; i < n; i++) {
        const float tmp = alpha * x[i];

        for (j = 0; j <= i; j++)
                ap[((i * (i+1)) / 2+j )] += x[j] * tmp;
}
```

Here it can be estimated that the required computation per data quantity is not enough to justify accelerator offload time. It is expected then that a GPU version of it is not going to have huge performance increments for small data.

2.2 GPU cuBLAS

This implementation was done directly reusing CUDA source code. NVIDIA CUDA [6] provides its own version of BLAS routines, which are heavily optimized to their architecture. Using the *cuBLAS* [8] library requires one call, an example is shown in Listing 4.

Listing 4. cuBLAS SSPR GPU Implementation

```
ret = cublasSspr(handle, mode, n, &alpha, cx, incx, cap);
if (ret != CUBLAS_STATUS_SUCCESS)
        err("cublasSspr: %d (%s)", ret, cublas2str[ret]);
```

This highly optimized code can be found in the package available to registered developers, inside `sspr.cu` and `sspr.h` files. The main routine is `cublasSspr()`. This implementation first loads into device shared memory elements reused during computation, then computes several matrix elements for hardware thread. It also uses cheap left bit-shifting instructions instead of expensive division-by-2 to locate elements inside the packed matrix.

The library documentation recommends the use of on utilities like: `cublasCreate()`, `cublasSetVector()`, `cublasGetVector()`, `cublasDestroy()`. They provide easier allocation of memory on the GPU device. The library also defines opaque data-types for parameters and error handling such as: `cudaError_t`, `cublasHandle_t`, `cublasStatus_t`, `cublasFillMode_t`.

[5] BLAS SSPR implementation.
[6] GSL SSPR implementation.

2.3 GPU Naive

A naive GPU implementation is useful to have a worst-estimate to be compared against potential optimizations. A direct translation to GPU using one thread per vector element to computing the result in parallel is shown in Listing 5. Here it is assumed that the number of elements in x it is close to the number of GPU threads.

Listing 5. Naive SSPR GPU Implementation

```
__global__ void sspr_naive_kernel(int uplo, int n, float alpha, const float *x, int incx, float *ap) {
        int i = blockIdx.x * blockDim.x + threadIdx.x;
        if (i < n) {
                const float tmp = alpha * x[i];
                int j = 0;
                for (j = 0; j <= i; j++)
                        ap[((i*(i+1))/ 2 + j)] += x[j] * tmp;
        }
}
```

How to execute the kernel is shown in Listing 6. CUDA will run a preprocessor transforming the code before performing actual compilation.

Listing 6. GPU Kernel execution

```
int threads = capabilities.maxThreadsPerBlock;
sspr_naive_kernel <<< (n / threads), (threads) >>> (uplo, n, alpha, cx, incx, cap);
```

2.4 GPU using shared-memory

The recommended approach to start optimizing GPU code is to use shared memory to reduce access time to data. Every thread on the same thread block loads in shared memory one element of the vector, this work is done in parallel and a barrier is used to synchronize. During computation, elements are then gathered from faster shared memory when possible, slower global memory is used otherwise. The implementation is shown in Listing 7.

Listing 7. Optimized SSPR GPU Implementation

```
__global__ void sspr_optimized_kernel(int uplo, int n, float alpha, const float *x, int incx, float *ap) {
        int i = blockIdx.x * blockDim.x + threadIdx.x;
        if (i < n) {
                int tid = threadIdx.x;
                extern __shared__ float cache[];
                float *xi = (float *) cache;
                xi[tid] = x[i];
                __syncthreads();
                const float tmp = alpha * x[i];
                int j = 0;
                for (j = 0; j <= i; j++) {
                        if (blockIdx.x * blockDim.x < j &&      blockIdx.x * blockDim.x + 1 > j)
                                ap[((i*(i+1))/ 2 + j)] += xi[j] * tmp;
                        else
                                ap[((i*(i+1))/ 2 + j)] += x[j] * tmp;
                }
        }
}
```

However, this method does not take into account any other potential optimization, it is included here to show that naive optimizations are not preferred and it is more useful to build upon CUDA implementation source code.

3 Performance Analysis

This section includes details on how the performance of the different implementations were conducted. The tools and procedure can be reused for any other BLAS kernel without major modifications.

3.1 GPU Registers

The CUDA `nvcc` compiler can be configured to show extra information about the number of registers used during execution. With this option the code can be optimized so register usage metric is minimal.

```
ptxas info: Compiling entry function 'sspr_naive' for 'sm_13'
ptxas info: Used 7 registers, 48+16 bytes smem, 4 bytes cmem[1]
ptxas info: Compiling entry function 'sspr_optimized' for 'sm_13'
ptxas info: Used 10 registers, 48+16 bytes smem, 4 bytes cmem[1]
```

It was discovered that running with the `-arch=compute_11` option did not included this output, but using `-arch=sm_13` instead solved the issue. The first one uses the compute capability version to specify the hardware target, while the second uses the hardware architecture generation.

3.2 Nvidia Visual Profiler

The `nvvp` tool depicts a visual execution profile of the program, useful when trying to understand where the computation is spending execution time. On our case, even with *cuBLAS* implementation the tool identified:

- there is no overlap between data transfer and computation
- the data transfer action does not fully saturate available memory bandwidth
- the computation does not fully load processing cores capacity

The affected GPU device does not have enough internal memory to run an interesting enough problem, the required computation per transfered bytes does not justify GPU offloading.

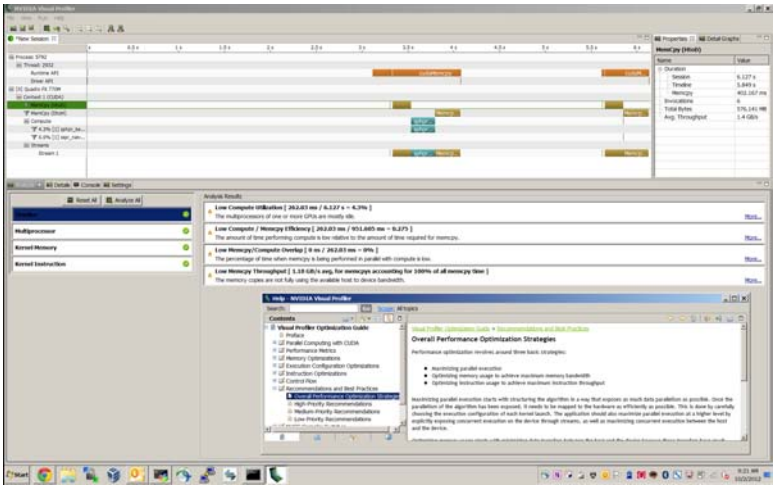


Fig. 3. GPU Profiler

3.3 Results

Figure 4 shows wall clock times in seconds with varying matrix sizes. [7] Here it can be seen that the CPU optimized version is taking the least time on all of the matrix sizes. On the other hand we double check that the GPU naive and shared-memory optimized versions are not a match against the CUDA one provided by cuBLAS. Here it can be clearly confirmed that the introduced shared memory optimization do not positively impact execution time, so *cuBLAS* strategy is hence superior.

[7] Times are the geometric mean of 32 executions in order to reduce measurement noise. To validate results the output matrix was reduced to a single figure being the common global sum of the elements.

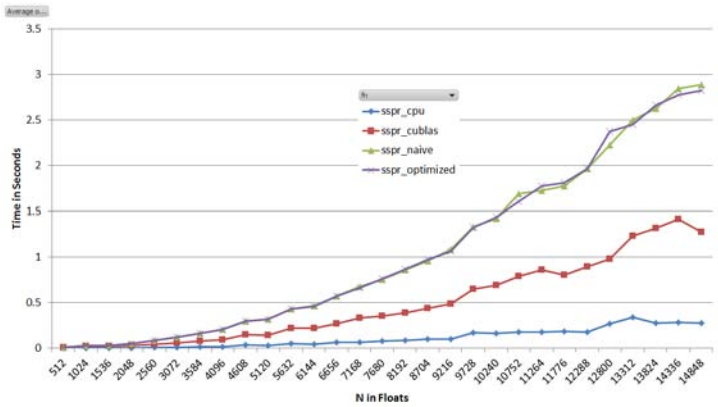


Fig. 4. Execution Time Speedup

3.4 Speedups

To measure speedups the input size was selected as big as possible to fit inside the available memory in GPU. This is always recommended to maximize computation per transfered byte. The time taken to transfer the data to the GPU is included on the measurements, the goal was to contrast the complete operation execution from start to end.

```
SSPR_N = 14848 floats (packed 110238976 floats)
SSPR_ALPHA = 3.141593
memory = 420 MB
cudaMemGetInfo.free = 462 MB
```

Most interesting speedup comparisons are shown in Table 2. The optimized CPU version has nearly 4x of cuBLAS version, and close to 8x of our naive implementations using GPU. It is interesting to note that cuBLAS optimization got 2x speedup when matched against our naive optimization with shared memory.

Table 2. Speedup comparisons

cublas (1.4995 seg)	cpu (0.389625 seg)	3.85x
naive (3.090625 seg)	cpu (0.389625 seg)	7.93x
optimized (2.97325 seg)	cpu (0.389625 seg)	7.63x
naive (3.090625 seg)	cublas (1.4995 seg)	2.06x
optimized (2.97325 seg)	cublas (1.4995 seg)	1.98x
optimized (2.97325 seg)	naive (3.090625 seg)	0.95x

3.5 Related Work

There is a related study conducted by Microsoft Research [9], that performed benchmarking of BLAS Level 2 routines in FPGA, CPU and GPU. Their findings in Figure 5 validates the results obtained as part of this work. An an optimized implementation in CPU is better than an optimized GPU implementation. Note that *Intel Math Kernel* library version is still far better than an optimized CPU version, as it uses advanced knowledge of the architecture of computing units inside the CPU.

It is worth to note that state-of-the-art GPU devices have increased their internal memory to cope with this limitation, up to 4GB in latest 2012 boards. If ECC is enabled to verify contents then this quantity is decreased 10%. A detailed analysis on how to better exploit GPU performance is reviewed in [10] using a complete application as case study.

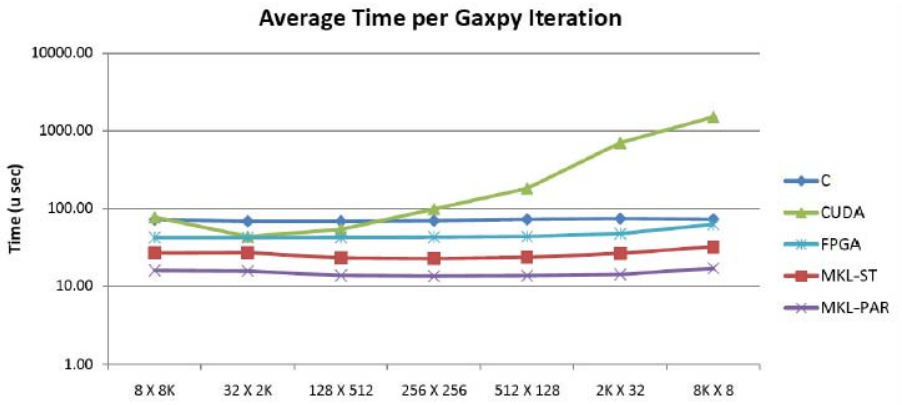


Fig. 5. Independent measurement of matrix-vector kernels (extract from [9])

4 Conclusions

This works provides a sample procedure to contrast BLAS kernel implementations after a experience with the SSPR operation. Source code pointers details are provided that guides through well-known implementation, also include performance analysis tools and their application example. In order to gather performance figures, it is always recommended to review the optimized code of BLAS instead of doing naive implementations. It is also worth to note that efficient GPU offload requires significant amounts of required computation per transfered byte. In this case, matrix-vector kernel computation showed that CPUs provide better results than GPUs; even when using highly optimized implementations of the kernel operation.

Regarding the experience with CUDA, the cuBLAS library documentation and interface properly support development, although they can still be improved. Some utility calls like `cublasAlloc()` are deprecated but still referenced by `cudaSspr()` and others, confusing the reader. The library does not provide a wrapper call that goes over the complete offload cycle: initialization, data transference to and from accelerator, explicit computation, etc. Using the documented primitives plus required error checking implies nearly 60 lines of code, just to offload one BLAS routine. cuBLAS also lacks an error formatter function to translate error codes to text representations (similar to `strerr()`). If it is required to support both Linux and Windows environments, the usual time keeping routines are not portable so a `gettimeofday()` stub was coded that could have been provided by the CUDA runtime instead.

Regarding further work, developing a framework to quickly benchmark compute kernels on different processing devices will be of value to domain experts researching what type of hardware to acquire. Ideally, including support for state-of-the-art BLAS implementations to provide figures from optimized algorithms. Also extending the comparison will be a good initial step, including other parallel programming techniques and technologies like OpenMP and MPI. Including results from other devices like FPGAs and co-processors would be another interesting option. There are other GPU optimizations that although require advanced knowledge (i.e. overlapping communication and computation, using texture memories) may result in better performance figures. Upcoming GPU architectures having more memory or new hardware-assisted features (i.e. zero-copy data replication) may also show different results.

References

1. Fan, Z., Qiu, F., Kaufman, A., Yoakum-Stover, S.: Gpu cluster for high performance computing. In: Proceedings of the 2004 ACM/IEEE conference on Supercomputing. SC '04, Washington, DC, USA, IEEE Computer Society (2004) 47–
2. http://gpgpu.org/: General-purpose computation on graphics hardware (april 2013)
3. Lawson, C.L., Hanson, R.J., Kincaid, D.R., Krogh, F.T.: Basic linear algebra subprograms for fortran usage. ACM Trans. Math. Softw. **5**(3) (September 1979) 308–323
4. Blackford, L.S., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., Heroux, M., Kaufman, L., Lumsdaine, A., Petitet, A., Pozo, R., Remington, K., Whaley, R.C.: An updated set of basic linear algebra subprograms (blas). ACM Transactions on Mathematical Software **28** (2001) 135–151
5. Golub, G.H., Van Loan, C.F.: Matrix computations (3rd ed.). Johns Hopkins University Press, Baltimore, MD, USA (1996)
6. NVIDIA Corporation: NVIDIA CUDA C Programming Guide v5.0. (October 2012)

7. Lee, V.W., Kim, C., Chhugani, J., Deisher, M., Kim, D., Nguyen, A.D., Satish, N., Smelyanskiy, M., Chennupaty, S., Hammarlund, P., Singhal, R., Dubey, P.: Debunking the 100x gpu vs. cpu myth: an evaluation of throughput computing on cpu and gpu. SIGARCH Comput. Archit. News **38**(3) (June 2010) 451–460
8. NVIDIA Corporation: NVIDIA CUBLAS Library v5.0. (October 2012)
9. Kestur, S., Davis, J.D., Williams, O.: Blas comparison on fpga, cpu and gpu. In: ISVLSI, IEEE Computer Society (2010) 288–293
10. Tomov, S., Nath, R., Ltaief, H., Dongarra, J.: Dense linear algebra solvers for multicore with gpu accelerators. In: Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on. (2010) 1–8

Mejoras en la eficiencia mediante *Hardware Locality* en la simulación distribuida de modelos orientados al individuo*

Silvana Lis Gallo [2, 3], Francisco Borges [1], Remo Suppi [1], Emilio Luque [1], Laura De Giusti [2], Marcelo Naiouf [2]

[1] Departamento de Arquitectura de Computadoras y Sistemas Operativos, Universitat Autònoma de Barcelona, Bellaterra, 08193, Barcelona, España.
[2] Instituto de Investigación en Informática LIDI (III-LIDI), Facultad de Informática, Universidad Nacional de La Plata, 50 y 120 2do piso, La Plata, Argentina.
[3] Becaria CONICET, Argentina.
{sgallo, ldgiusti, mnaiouf}@lidi.info.unlp.edu.ar
{Remo.Suppi, Emilio.Luque}@uab.cat, francisco.borges@caos.uab.cat

Resumen. La simulación de altas prestaciones aplicada a modelos orientados al individuo es de gran interés en la comunidad científica por la precisión que aportan sus datos, pero por contrapartida necesita grandes capacidades cómputo. Por ello, es necesaria la utilización de técnicas y métodos que permitan aprovechar toda la potencia de cómputo disponible para obtener el máximo *speedup* y eficiencia sobre la arquitectura. Por otro lado, el incremento del número de *cores*, cachés compartidas y memoria de los nodos ha introducido una complejidad en la arquitectura que puede afectar seriamente a las prestaciones/eficiencia de las aplicaciones si no se hace una distribución correcta teniendo en cuenta la jerarquía subyacente. El presente trabajo muestra las mejoras introducidas en la simulación distribuida de un modelo orientado al individuo (*Fishschools*) en sistemas *multicores* utilizando *hardware locality* (hwloc), la cual provee información sobre los procesadores de los nodos. Esta información será utilizada por la aplicación para adaptar las estrategias de ubicación de los procesos dependiendo de la afinidad hardware.

Palabras Clave: Simulación Paralela y Distribuida. Modelos Orientados a Individuos. Simulación de eventos discretos. Cluster de multicore. Evaluación de prestaciones.

1 Introducción

El estudio de la dinámica de poblaciones es un área de gran interés en el ámbito académico y ha sido el objetivo fundamental en el desarrollo de modelos biomatemáticos y una herramienta necesaria en la ecología demográfica para analizar y cuantificar las variaciones que sufren ciertas poblaciones a través del tiempo. Existen dos enfoques con los cuales se puede modelar la dinámica de poblaciones:

* El presente trabajo ha sido financiado por el proyecto del MICINN-España TIN2007-64974 y MINECO-España TIN2011-24384.

modelado basado en ecuaciones y modelado orientado a individuos. En los modelos basados en ecuaciones, las propiedades del sistema se obtienen mediante la resolución de un sistema de ecuaciones generalmente diferenciales y se obtienen resultados globales para el conjunto y con un alto grado de abstracción. En los modelos orientados a individuos, las propiedades del sistema emergen como resultado de la interacción y del comportamiento de los individuos del sistema y puede proporcionar resultados más cercanos a la realidad y con un alto grado de detalle. Estos modelos son de alta complejidad por lo cual no es posible su resolución analítica y deben hacerse por medio de simulación computacional.

Por otro lado, la aparición de arquitecturas distribuidas y procesadores con varios núcleos ha permitido el desarrollo de modelos complejos y a gran escala utilizando técnicas de simulación distribuida para reducir los tiempos de ejecución y obtener resultados en tiempo aceptables. No obstante, el aumento del número de procesadores interconectados a través de una red permite ejecutar gran cantidad de procesos, pero se debe tener especial cuidado en el impacto de las comunicaciones. Rápidamente se pueden degradar las prestaciones de la aplicación ya que el tiempo de comunicación puede ser más elevado que el cómputo paralelo y que en consecuencia no haya beneficios reales en su ejecución distribuida.

En este sentido, se han desarrollado técnicas y métodos que permiten simular modelos orientados al individuo de gran escala con el fin de disminuir el tiempo total de ejecución. La investigación realizada en [3] tiene por objetivo la simulación distribuida de *Fishschools*, un modelo orientado al individuo complejo, en el cual se demuestran las posibilidades de la simulación distribuida aplicando algoritmos de simulación conservativos. El trabajo realizado con este tipo de modelos tiene diferentes puntos que pueden comprometer las prestaciones y la escalabilidad del modelo: la asignación de individuos a los nodos de cómputo, cómo se agrupan para reducir la complejidad del algoritmo y el balanceo dinámico de carga o la importancia de cómo se realizan las comunicaciones entre los procesos distribuidos [5, 6, 16]. Todos los experimentos realizados demuestran que el simulador logra valores de *speedup* muy buenos para este tipo de modelos y que el modelo de simulación es escalable.

Sin embargo, ¿existe la posibilidad de mejorar la eficiencia del simulador y aprovechar de forma más adecuada la arquitectura subyacente?. Sería posible mejorando uno de los aspectos importantes y no considerados hasta el momento: la complejidad de la topología del hardware y la forma en que se realiza la distribución de los procesos sobre los *cores* de la arquitectura y cómo afectan las diferentes jerarquías de memoria a la ejecución de procesos. El objetivo del presente trabajo es analizar cómo el sistema operativo realiza la asignación de procesos a los nodos de cómputo-*cores* y cómo esto afecta a las comunicaciones (OpenMPI) entre los diferentes cores-nodos para obtener así la mejor estrategia con el fin de mejorar la eficiencia de la simulación bajo estudio utilizando *hardware locality* [18] (API que se describe en la sección 3 y cuya aplicación permite realizar las asignaciones de procesos de la simulación distribuida a los diferentes *cores*).

Este artículo se organiza con la siguiente estructura. En la sección 2 se presentan los modelos orientados a individuos, el simulador de banco de peces original y un resumen de las características anteriormente implementadas en el mismo. La sección 3 detalla las estrategias para utilizar la información de la arquitectura para la selección

y asignación de procesos a *cores* y en la sección 4, se describen las pruebas realizadas y resultados obtenidos en las mismas. Por último, en la sección 5, se presentan las conclusiones y trabajos futuros.

2 Modelos orientados al individuo

Los modelos orientados al individuo (IoM) permiten entender la dinámica del comportamiento de un sistema y consisten en una cantidad fija de individuos autónomos, para los cuales se definen reglas de interacción y atributos individuales que se mantienen a lo largo del tiempo. Además, incorporan un entorno donde ocurren las diferentes interacciones y un conjunto de parámetros que permite modelar estas interacciones y su movimiento en un espacio tridimensional.

Estos modelos también permiten incluir diferentes tipos de individuos dentro del mismo modelo con diferentes reglas de comportamiento y distintos valores de los atributos para modelar la interacción entre especies, su vinculación con el entorno, y también permite incluir diferentes mecanismos de aprendizaje, energía, obstáculos, roles (presa-depredador), etc. Algunos modelos orientados a individuos son también espacialmente explícitos, como es el caso de las simulaciones en que los individuos son asociados a una ubicación en el espacio geométrico y que también pueden mostrar patrones de movimiento (por ejemplo, los individuos pueden cambiar su posición relativa en un espacio geométrico). En la literatura existen diversos estudios de simulaciones espacialmente explícitas orientadas a individuos, como es el caso de aves [2,7], insectos [4,8,9], mamíferos [10] y peces [11-15].

2.1 Fishschools y la simulación distribuida

El presente trabajo utiliza como IoM a *Fishschools* que es un modelo biológico ampliamente validado y que representa el comportamiento de un conjunto de peces que es considerado como uno de los grupos sociales más frecuentes en el mundo animal [11,1]. Esta agrupación social muestra propiedades emergentes complejas, como por ejemplo: una fuerte cohesión de grupo (se mantiene la formación de grupo a través del tiempo), y un alto nivel de sincronización (los peces se mantienen nadando hacia la misma dirección y con la misma rapidez).

Para la simulación de este modelo se ha desarrollado un simulador distribuido que implementa algoritmos conservadores para realizar la sincronización entre procesos lógicos [3], por lo que éstos se bloquearán hasta que el procesamiento sea seguro. Para describir el comportamiento de los peces, el modelo considera que cada pez cambia su posición y orientación en pasos discretos de tiempo (time-driven) y los nuevos valores dependen de la posición y orientación de un número fijo de vecinos cercanos que estarán en función de la visión de pez. La influencia de los vecinos para un individuo en particular depende de su posición tiempo-espacio y la selección de la influencia de los vecinos está considerada en tres áreas de visión: atracción, repulsión y orientación paralela, como se indica en la Figura 1.

Dependiendo de la posición espacial y temporal de sus vecinos, el pez elige entre los tres patrones de comportamiento: **repulsión** - para evitar la colisión entre peces

del mismo grupo (cambiando la orientación del ángulo mínimo de rotación, logrando que la orientación del pez y la orientación de su vecino sean perpendiculares, Figura 2a), **orientación paralela** - el grupo se mueve en la misma dirección (haciendo coincidir la orientación del pez con la orientación de sus vecinos, Figura 2b), y **atracción** - para mantener la cohesión del grupo (dirigiendo su orientación hacia la posición de su vecino, la Figura 2c).

Figura 1. Áreas de visión

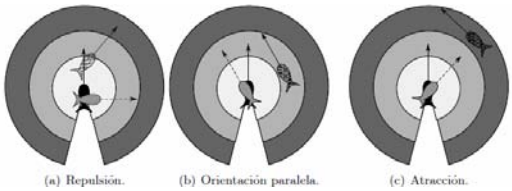


Figura 2. Patrones de comportamiento.

La orientación final del individuo surge a partir de la aplicación ponderada de las acciones resultantes con cada vecino y en el espacio tridimensional. Es importante notar que el modelo de simulación debe realizar este cálculo por cada individuo y para cada paso de simulación lo cual se deriva en altas necesidades de cómputo cuando se trabajan con un número importante de individuos (de 64k a 512k individuos en el caso tratado).

Con el fin de mejorar el rendimiento en la simulación, los autores muestran en [3] una importante contribución a la generalidad del modelo original -en dos dimensiones- [1] añadiendo la interacción depredador-presa, evasión de obstáculos cilíndrica, y una representación interactiva de los límites del mundo en el modelo 3D. Este modelo si bien es más complejo también produce resultados más cercanos a la realidad reduciendo el grado de abstracción y proporcionando buena escalabilidad y rendimiento cerca de los valores de *speedup* ideales.

En [5] se presenta un enfoque de agrupación en la distribución de los individuos en la arquitectura con el fin de obtener el mejor rendimiento de la simulación ya que el desbalance de carga es un problema importante en la simulación distribuida. El enfoque basado en clústeres consiste en determinar el conjunto de individuos óptimo a cada elemento de cómputo de forma tal que permita reducir la interacción entre los individuos que están lejos. En [6] considerando el problema de la distribución se presenta una nueva estrategia que complementa a la anterior y que incluye el equilibrio de carga dinámico para evitar que después de una asignación inicial óptima en los nodos de cómputo se tengan desbalances debido al movimiento de los individuos y el paso del tiempo de simulación. Esta estrategia de balanceo de carga se basa en reconfigurar y redistribuir la carga de trabajo local haciendo nuevas agrupaciones y migrando los individuos desde un nodo de cómputo hacia otro.

Finalmente, dado que la comunicación es uno de los puntos débiles de la simulación distribuida en [17], se comparan tres estrategias de comunicación implementadas en el simulador distribuido: comunicación asíncrona y síncrona de paso de mensajes y *bulk-synchronous parallel*. En este trabajo se demuestra que las simulaciones *time-driven* no siempre aumentan el rendimiento mediante el uso de estrategias de comunicación sincrónica. Los resultados demuestran que la comunicación sincrónica obtiene peores resultados en términos de tiempos de ejecución en comparación con la estrategia de comunicación asincrónica.

Una vez optimizados los aspectos del modelo, asignación y balanceo de carga de los individuos, y prestaciones de las comunicaciones queda por analizar y optimizar la eficiencia y uso de los recursos de la arquitectura, Es por ello que en el presente trabajo se estudia el impacto de la afinidad de procesos dinámica utilizando *Hardware Locality* [18] en el simulador distribuido de modelos orientados al individuo con el objetivo de analizar cuándo existen mejoras en las prestaciones y en la eficiencia en las arquitecturas de cómputo actuales.

3 *Hardware Locality* aplicada a la simulación distribuida de *Fishschool*.

Los procesadores *multicores* son ampliamente utilizados en la computación de altas prestaciones y es una tendencia donde la arquitectura es cada vez de mayor complejidad y donde adquiere especial relevancia el acceso no uniforme a la memoria de los nodos de cómputo. Es por ello que se requiere una asignación ordenada y en base a estrategias predefinidas de procesos y datos en función de su afinidad si se desea aprovechar al máximo las posibilidades de la arquitectura subyacente. Por otro lado hay que tener en cuenta que en grandes infraestructuras, generalmente controladas por sistemas de colas, las políticas de asignación de las colas pueden ser totalmente contrarias a las necesidades de la aplicación distribuida y el usuario debe disponer de herramientas que bajo unas situaciones predefinidas pueda asignar sus procesos de la forma más conveniente (y en forma dinámica) para la ejecución de la aplicación.

También es necesario tener en cuenta que la utilización de hardware de red puede resultar de acceso no uniforme ya que las tarjetas de interfaz pueden estar más cerca de algunos procesadores que de otros. Es por ello que se debe analizar estos aspectos pues pueden tener un impacto considerable en el rendimiento de las comunicaciones y afectar a las aplicaciones de paso de mensajes, como por ejemplo las que utilizan OpenMPI.

Una posible estrategia de trabajo es actuar sobre la colocación (*mapping*) de los procesos orientados a cómputo y de los orientados a comunicaciones para que de esta forma se pueda mejorar las prestaciones de la aplicación facilitando tanto el acceso a datos como el acceso a las interfaces de red para aquellos procesos que lo necesiten de forma intensiva.

Con el fin de poder explotar el potencial de las arquitecturas *multicores* para el simulador distribuido bajo estudio se ha utilizado la API hwloc [18] que permite realizar una abstracción de la topología hardware al desarrollador de la aplicación. Asimismo, estas características permiten en tiempo de ejecución realizar la asignación controlada de recursos (*cores*) a los procesos lógicos de la aplicación de acuerdo a la estrategia preseleccionada y el *hardware* subyacente donde se ejecutarán dichos procesos.

Normalmente, será el sistema operativo quien elegirá los núcleos donde se crearán los procesos MPI y estos podrán ser desplazados, de acuerdo a sus políticas de optimización, de un núcleo hacia otro como se muestra en la Figura 3 sobre 4 procesos MPI y su asignación a un procesador de 8 *cores* y dos *sockets*.

Sin embargo, esto puede limitar el rendimiento paralelo de la aplicación en el contexto de HPC ya que se pierden las ventajas de la memoria caché y la localidad cuando el proceso migra de un núcleo a otro. Como se puede observar, los procesos se han creado bajo un criterio predefinido por el SO con el que no compartirán la cache L2 y por lo cual no podrán compartir información entre los procesos. En cambio si se mantiene un proceso "unido" a un *core* de un núcleo y se evita el cambio de contexto tendrá como resultado una mejor utilización de caché y producirá una mejora de las comunicaciones entre procesos [16].

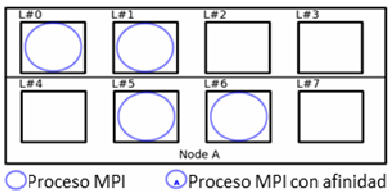


Figura 3. Asignación por el SO de cuatros procesos MPI

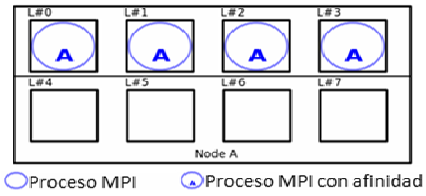


Figura 4. Cuatro procesos MPI asignado a cuatro cores por Hwloc

Para evitar estos problemas, cada proceso de MPI debería se asignado a un núcleo de forma consecutiva, de acuerdo con el número del MPI Rank. De esta forma, los procesos adyacentes estarán espacialmente cercanos y podrán tomar ventaja de L1 y L2 (Figura 4).

La estructura de datos utilizada en el simulador distribuido se basa en una lista de radio fijo de clústeres [5] (Figura 5), que se mantiene en cada proceso lógico de simulación (LP) y donde el proceso dispone de toda la información para realizar el siguiente paso de simulación. Es importante analizar que la forma de acceso a los datos y su asignación a elementos de cómputo puede favorecer las prestaciones teniendo en cuenta cómo se almacenan los individuos en un nodo y la forma en que se defina la afinidad.

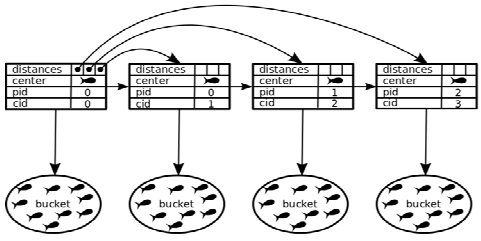


Figura 5. Estructura de datos que mantiene cada LP.

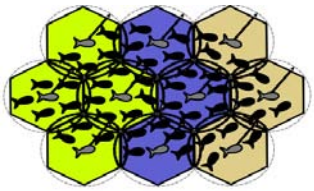


Figura 6. Áreas de clústeres adyacentes (mismo color) deberían ser asignadas a *cores* adyacentes.

Esta estructura de datos permite definir áreas en las que los individuos pueden interactuar sólo con los que pertenecen a las zonas adyacentes y ello permite reducir el tiempo de cálculo de los individuos que participan en el mecanismo de selección vecinos (Figura 6). Si estas áreas son asignadas a elementos que puedan compartir la caché del *socket* los procedimientos que utilizan datos de su área y de áreas vecinas (método de descubrimiento de vecinos) tendrán acceso a los datos de forma más eficiente que si se asignan a *cores* que no compartan esta memoria.

Para realizar la experimentación y analizar el impacto de la afinidad en procesos MPI, el código del simulador ha sido modificado para que permita obtener

conclusiones de cómo afecta la afinidad a la simulación distribuida de modelos orientados al individuo sobre una arquitectura *multicores*.

4 Pruebas realizadas

Para la experimentación se ha utilizado una arquitectura de tipo Blade de 12 bahías y donde cada *blade* dispone de 2 procesadores quad core Intel Xeón (e5405@2.0GHz) con una caché L1 privada de 64Kb (dividida en 32Kb para instrucciones y otros 32Kb para datos), y cache L2 de 2 x 6Mb entre par de núcleos. En relación a la memoria principal (RAM), 8 de los *blades* disponen de 10Gb compartida entre ambos procesadores y los 4 nodos restantes tienen 2 Gb en la misma configuración. La Figura 7 muestra la imagen generada por el comando *lstopo* de hwlock1.6.1.

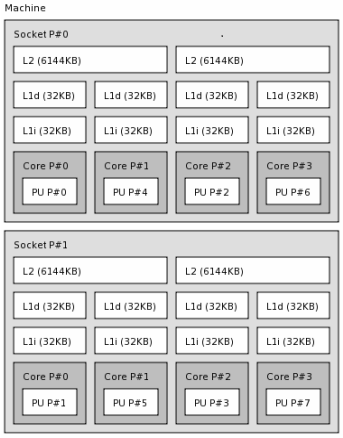


Figura 7. Arquitectura del *blade* obtenida por el comando *lstopo* (hwloc).

Para analizar el comportamiento de cada solución paralela implementada, se midió tiempo de cómputo y comunicaciones así como también se calcularon los valores *speedup*, y eficiencia para cada escenario. Por un lado, se realizaron pruebas con hasta 64 *cores* de la arquitectura en nodos con igual cantidad de memoria (10Gb). Los distintos escenarios de prueba se realizaron con la misma cantidad de ciclos de simulación (250 pasos), y escalando el tamaño de la población de individuos (8K, 16K, 32K, 64K, 128K individuos), y las diferentes cantidades de *cores* utilizados (en total 4, 8, 16, 32 y 64). Por otro lado, para analizar el impacto de la heterogeneidad en la solución, se realizaron pruebas con 64 y 96 nodos activando y desactivando el uso de *hwloc*.

4.1 Pruebas en el entorno homogéneo

Los valores obtenidos para los tiempos de cómputo y comunicación para cada uno de los escenarios son los visualizados en la Figura 8. Como se puede observar, en cuanto a los tiempos de cómputo (Figura 8a), el uso de *hwloc* logra una disminución para todos los escenarios de prueba. Por otro lado, cuando el escenario de prueba es grande (128K individuos), el tiempo de comunicación (Figura 8b), presenta una mayor

reducción que en los demás escenarios de prueba. Esto se debe a que la ejecución de la simulación para esta cantidad de *cores* implica una gran cantidad de mensajes, por lo que permite que el uso de la API *hwloc* sea más notable.

En la Figura 9, se muestran *speedup* y eficiencia obtenidos en las simulaciones. En todos los casos, se observa una mejora que alcanza 2%, siendo el mejor resultado el obtenido por el escenario 4c-8k, lo cual es esperable, porque cuando se incrementa el número de procesos/*cores*, la eficiencia disminuye ya que los mensajes y el número de procesos hace que los tiempos de espera por comunicación adquieran relevancia. Es importante notar que la distribución de los procesos (manual a través *hostfile*) no es la peor (ni aleatoria como podría ser en un sistemas de colas) ya que el algoritmo de asignación de sistema operativo está haciendo una asignación similar a la de hwloc.

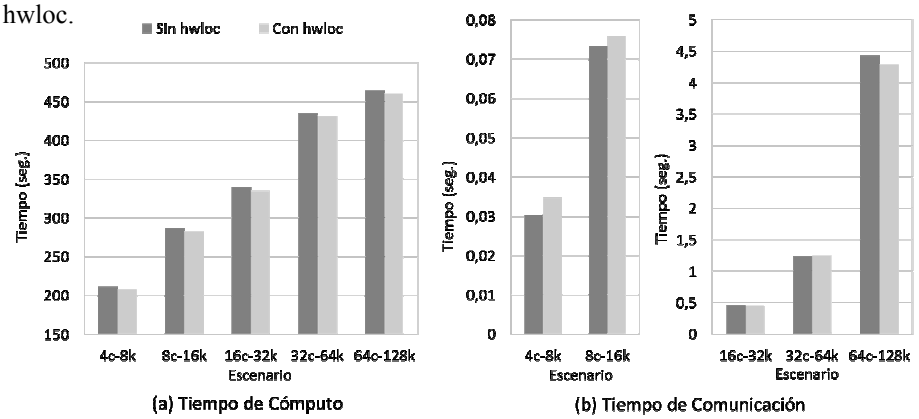


Figura 8. Tiempos de cómputo y comunicaciones para los diferentes escenarios.

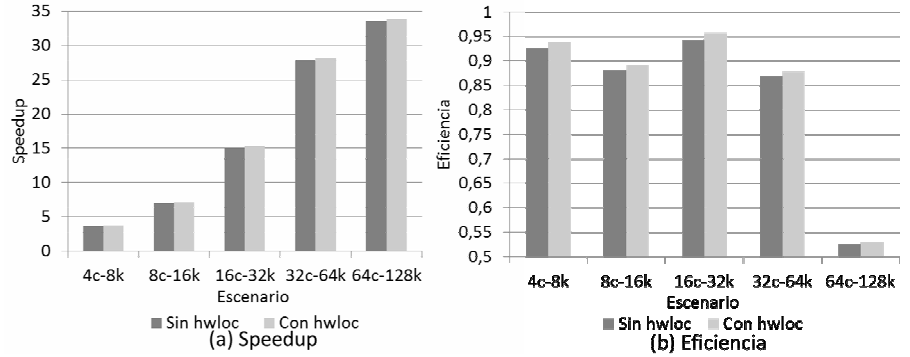


Figura 9. Speedup y eficiencia para las pruebas homogéneas.

Además, se debe considerar que las herramientas como OpenMPI se encuentran optimizadas para ambientes *multicore* (haciendo intercambios de punteros entre buffers de mensajes), por lo que, definir procesos en la misma y asignarlos manualmente (mediante un *hostfile)* permite realizar una buena distribución, que se ha visto mejorada con el uso de *hwloc*. Este resultado permite pensar que la incorporación de herramientas de programación de memoria compartida (como por ejemplo OpenMP) podría resultar de gran ayuda para el perfeccionamiento de las soluciones sobre clúster de *multicore*.

4.2 Pruebas en el entorno heterogéneo

En este caso se utilizaron las hojas de 2Gb y 10Gb intercaladas para la ejecución, con un escenario de 128K individuos, con 64 y 96 cores, permitiendo obtener los valores de eficiencia de la Figura 10. Como se observa, si el mismo caso de prueba se ejecuta con y sin *hwloc* en los distintos entornos (Figura 10a), se consigue mejor eficiencia en una arquitectura homogénea. Por otro lado, cuando se utiliza *hwloc* la eficiencia es levemente mayor, lo cual indica que esta solución es favorable en ambos entornos, y mayormente en el entorno heterogéneo. Al ser los incrementos de la eficiencia muy pequeños, se refuerza la necesidad de buscar una alternativa que optimice el trabajo dentro de cada nodo de acuerdo a su arquitectura (en especial L1 y L2).

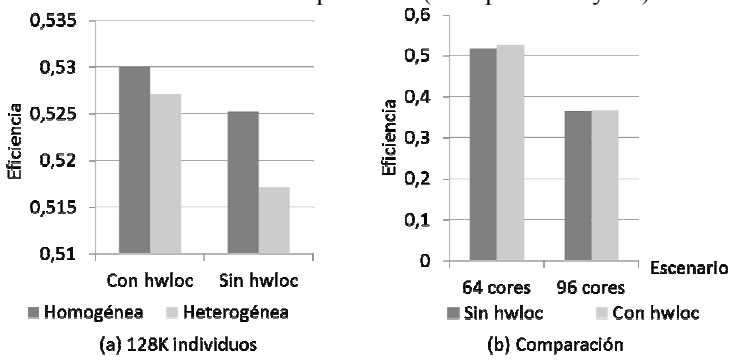


Figura 10. Eficiencia en soluciones con heterogeneidad de arquitectura

Por último, si se compara la eficiencia obtenida para un mismo caso de prueba (128K individuos), para un mayor número de procesadores en un ambiente heterogéneo (como es el caso de 96 *cores* en la Figura 10b), se puede observar que el uso de *hwloc* sigue siendo levemente favorable para la ejecución en el entorno heterogéneo.

5 Conclusiones y trabajos futuros

Como resultado del presente trabajo se obtuvo una reducida mejora que, a medida que el número de *cores* es más elevado, tiende a disminuir por el propio efecto de los mensajes. Además, considerando que la ejecución de las pruebas fue en un entorno sin manejo de colas, el SO realiza la asignación de manera similar a *hwloc*. Así mismo es importante notar que el simulador mantiene su escalabilidad y hwloc no perjudica a valores elevados de *cores*/procesos. En el sentido de los incrementos de prestaciones será necesario explorar nuevas alternativas que permitan reducir la cantidad de mensajes de la aplicación, como puede ser el uso de herramientas de programación sobre memoria compartida (por ejemplo OpenMP). Por otro lado, si bien el simulador hace un balance de carga, no lleva a cabo un balance de cómputo, que perjudica los resultados obtenidos, convirtiendo este aspecto en otro ítem a optimizar.

Agradecimientos

Los autores desean agradecer al Dr. Roberto Solar las sugerencias al presente trabajo así como las aportaciones e investigaciones realizadas en las versiones anteriores del simulador distribuido. El presente trabajo ha sido financiado por el proyecto MICINN-España TIN2007-64974 y proyecto MINECO-España TIN2011-24384.

Referencias

1. A. Huth, C.Wissel, The simulation of fish schools in comparison with experimental data, Ecological Modelling, State-of-the-Art in Ecological Modelling, Proceedings of ISEM's 8th International Conference.75-76 (1994) 135–146.
2. C. W. Reynolds, Flocks, herds and schools: A distributed behavioral model, SIGGRAPH Computer Graphics. 21 (1987) 25–34.
3. R. Solar, R. Suppi, E. Luque, High performance individual-oriented simulation using complex models, Procedia Computer Science 1 (1) (2010) 447 – 456.
4. X. Hu, Y. Sun, Agent-based modeling and simulation of wildland fire suppression, in: Proceedings of the 39th conference on Winter simulation, IEEE Press, USA, (2007) 1275–1283.
5. R. Solar, R. Suppi, E. Luque, High performance distributed cluster-based individual-oriented fish school simulation., Procedia CS 4 (2011) 76–85.
6. R. Solar, R. Suppi, E. Luque, Proximity load balancing for distributed cluster-based individual-oriented fish school simulations, Procedia Computer Science 9 (0) (2012) 328 – 337.
7. R. O. Saber, R. M. Murray, Flocking with obstacle avoidance: cooperation with limited communication in mobile networks, Proceedings 42nd IEEE Conference on Decision and Control, 2003, Vol. 2, (2003) 2022–2028.
8. E. Bonabeau, M. Dorigo, G. Theraulaz, Swarm intelligence: from natural to artificial systems, Oxford University Press, USA, (1999).
9. J. Kennedy, R. C. Eberhart, Swarm intelligence, Morgan Kaufmann Publishers, USA, 2001.
10. S. Gueron, S. Levin, D. Rubenstein, The dynamics of herds: From individuals to aggregations, Journal of Theoretical Biology 182 (1996) 85–98.
11. A. Huth, C. Wissel, The simulation of the movement of fish schools, Journal of Theoretical Biology 156 (3) (1992) 365 – 385.
12. I. Aoki, A simulation study on the schooling mechanism in fish, Journal of the Japanese Society of Scientific Fisheries 48 (8) (1982) 1081–1088.
13. R. Vabø, G. Skaret, Emerging school structures and collective dynamics in spawning herring: A simulation study, Ecological Modelling 214 (2-4) (2008) 125–140.
14. J. K. Parrish, S. V. Viscido, D. Grnbaum, Self-organized fish schools: An examination of emergent properties, Biological Bulletin 202 (2002) 296–305.
15. J. C. Gonzlez, C. Dalforno, R. Suppi, E. Luque, A fuzzy logic fish school model, Lecture Notes in Computer Science, Vol. 5544 (2009) 13–22.
16. Open MPI Team. FAQ: General run-time tuning., Feb 2012. Visited on July 23, 2013.
17. R. Solar, F. Borges, R. Suppi, and E. Luque. Improving Communication Patterns for Distributed Cluster-based Individual-oriented Fish School Simulations. Procedia Computer Science, 18(0), (2013) 702-711.
18. François Broquedis, Jérôme Clet-Ortega, Stéphanie Moreaud, Nathalie Furmento, Brice Goglin, Guillaume Mercier, Samuel Thibault, and Raymond Namyst. hwloc: a Generic Framework for Managing Hardware Affinities in HPC Applications. 18th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, IEEE Computer Society Press, (2010) 180-186. DOI: 10.1109/PDP.2010.67.

Un método de sintonización para mejorar la salida de un modelo computacional de cuenca de ríos

Adriana Gaudiani[1,3], Emilio Luque[2], Armando Di Giusti[3], and Marcelo Naiouf[3]

[1] Instituto de Ciencias, Universidad Nacional de General Sarmiento, Los Polvorines, Argentina
agaudi@ungs.edu.ar
[2] Dept. de Arquitectura de Computadores y Sistemas Operativos, Universitat Autònoma de Barcelona, 08193 Bellaterra(Barcelona)España
[3] Instituto de Investigación en Informática LIDI (III-LIDI), Universidad Nacional de La Plata, Bs. As., Argentina

Abstract. Los modelos computacionales que simulan fenómenos naturales se pueden comportar de manera muy próxima a la real, pero debido a múltiples factores los resultados simulados difieren de los resultados reales. Una fuente de error es la falta de certeza en los valores de los parámetros de entrada. Este trabajo constituye un primer paso en el enunciado de una metodología que busca mejorar la capacidad de predicción de un simulador, aplicado a un modelo computacional de cuenca de ríos y, en particular, utilizando el modelo del cauce del Río Paraná. Se presenta un método computacional para la sintonización de los valores de los parámetros de entrada de dicho modelo, con el objetivo de minimizar el error entre la salida del simulador y la realidad observada. Este proceso de sintonización se lleva adelante aplicando una técnica de simulación paramétrica, la cual conlleva a ejecutar un gran número de simulaciones haciendo necesario utilizar recursos de cómputo de alto rendimiento y técnicas de paralelización.

1 Introducción

La modelización y la simulación de inundaciones provocadas por el desborde de ríos brindan modelos computacionales para el estudio y la predicción de estos fenómenos naturales con el objetivo de estudiar, simular y predecir su comportamiento.

El ingrediente esencial de cada modelo son las variables y los parámetros. Las variables son cantidades físicas y los parámetros controlan el comportamiento de las variables. Modelizar sistemas de la naturaleza, que son sistemas reales complejos, implica normalmente el uso de muchos parámetros y variables de entrada[1]. Al modelizar y simular el flujo y desborde de ríos, se ingresan los valores de los parámetros de entrada a un simulador computacional, siendo la salida del modelo, hidrogramas de caudal e información sobre el evento de inundación. Aunque los modelos utilizados consideren la mayor cantidad posible de variables involucradas en el proceso, tratando de simular de la manera más

certera el fenómeno, existen otros factores que no permiten obtener resultados confiables. Por diversos motivos, los valores de los datos de entrada son imprecisos provocando diferencias entre los resultados dados por el simulador y los medidos en la realidad [2]. Se detallan algunas de estas causas a continuación.

Los parámetros de entrada a la simulación son medidos en unas pocas estaciones distribuidas a lo largo del cauce del río, siendo necesario interpolar sus valores en todo el dominio. Algunos datos, como las precipitaciones, cambian dinámicamente durante todo el proceso y otros son magnitudes físicas que arrastran errores de medición o deben medirse de manera indirecta (coeficiente de Manning, altura de albardones, conductancia, etc.). Los datos de salida son la altura del cauce del río, calculada en estaciones a lo largo de su recorrido y en sucesivos intervalos de tiempo. La idea principal de la investigación es minimizar el impacto en los datos de salida provocado por la incertidumbre en los valores de los parámetros de entrada al simulador, para brindar una mejora que ayude a los ingenieros que trabajan con cuencas hidráulicas a producir alertas a la población ante eventos de desborde del cauce de los ríos.

Una manera de abordar el problema de la incertidumbre es mediante una fase de ajuste de parámetros y posteriormente, la fase de verificación de la mejora y de su impacto en la capacidad predictiva del simulador[8]. La etapa de ajuste se realiza para obtener un conjunto de parámetros que minimice las diferencias entre la salida simulada y la real. En esta etapa se aplica la técnica de simulación paramétrica para procesar una gran cantidad de escenarios (cada configuración de parámetros del sistema simulado) y calcular una medida del ajuste para cada uno.

El método presentado es propio del mundo de las altas prestaciones. La sintonización es posible con la ejecución de una enorme cantidad de escenarios, consumiendo un elevado tiempo de ejecución y requiriendo el uso de técnicas de cómputo paralelo a nivel básico para lanzar la mayor cantidad de ejecuciones posibles de manera simultánea.

La presentación está organizada de la siguiente manera: En el capítulo 2 se presentan las características del simulador utilizado y el modelo del Río Paraná con el cual se lleva adelante este trabajo. En el capítulo 3, se presenta la metodología propuesta para el método de sintonización del simulador implementada sobre una muestra acotada de escenarios. En los capítulos 4 y 5 se detallan las experiencias realizadas con su análisis y en el capítulo 6 las conclusiones y trabajo futuro.

2 El simulador EZEIZA

Se seleccionó un programa de simulación utilizado actualmente para brindar alertas ante posibles eventos de inundaciones, el cual es un simulador del cauce y flujo de agua en ríos. El software seleccionado es Ezeiza [7] cuyas características se brindan a continuación.

2.1 Modelo computacional para cauce de ríos: Ezeiza

El software de simulación utilizado, Ezeiza V.6, fue desarrollado en el Laboratorio de Hidráulica Computacional del Instituto Nacional del Agua (INA)[4]. Ezeiza es un modelo computacional para el cálculo de la traslación de ondas en ríos y canales que comenzó a desarrollarse en la década del '70. Se basa en un análisis unidimensional, expresado matemáticamente mediante las ecuaciones de Saint Venant(1891) y resolviendo las mismas mediante técnicas numéricas [4]. Como condiciones iniciales, deben proveerse las distribuciones de nivel y caudal sobre todo el sistema. Su versión más actual permite el tratamiento de una red de flujo arbitraria. Actualmente, es usado con un modelo hidrodinámico del Río Paraná y se utiliza como herramienta adicional para el pronóstico de crecidas y bajantes de la Cuenca del Plata, tarea que está llevando a cabo el Servicio de Información y Alerta Hidrológico (SIyAH) del INA. Cabe destacar que el Río Paraná recorre una de las áreas más pobladas e industrializadas de Sudamérica en el tramo modelado, requiriendo implementar tecnologías que mejoren constantemente los pronósticos.

La validación del modelo fue efectuada por los ingenieros del INA. En el informe de la validación y calibración del simulador, realizado por el Ing. Menéndez en 1996, ya se expresaba la necesidad de mejorar la precisión de los resultados simulados. Este análisis se retoma en un exhaustivo estudio de rendimiento realizado en 2011 por el Ing. Latessa para el INA [6].

La elección de este simulador se hizo considerando que:

- Exporta los resultados en archivos que pueden ser tratados desde paquetes estadísticos y/o matemáticos.
- Permite realizar simulaciones paramétricas modificando los valores de los parámetros en archivos de entrada, situación fundamental para el proceso de sintonización.

El estudio de Latessa, tuvo como objetivo mejorar la modelización del río Paraná y brindar más certeza a los datos de entrada, debido a que se ingresaban parámetros medidos décadas atrás. Se mejoró mucho la precisión del modelo,

[4] $http://www.ina.gov.ar/lha/index.php?lha=38$

como puede verse en dicho informe. Aunque, aun se pueden detectar diferencias entre los valores reales y los datos de salida del simulador en las estaciones de seguimiento. Estas diferencias son las que se intentan minimizar en este trabajo, haciendo un estudio del error de predicción del simulador.

2.2 Características del modelo del río Paraná

El modelo hidrodinámico del Río Paraná y Paraguay, que corre sobre Ezeiza, fue diseñado por el INA para simular su comportamiento en los tramos que van desde la represa de Yacyretá (Corrientes) hasta Villa Constitución (Santa Fe). Los datos de la red de cálculo pueden verse en la Tabla 1.

Tabla 1. Red de Cálculo: Modelo del Paraná

$Filamento$	$Curso$	$Long.(Km)$	$Secciones$	$Bordes\ Ag.Arr.$	$Bordes\ Ag.Aba.$
1	Paraná	1.083	76	Caudal Q	Nivel H
2	Paraguay	376	77	Nivel H	——-

Los parámetros críticos del modelo son el coeficiente de Manning y el nivel de los albardones. El coeficiente de fricción de Manning representa una cuantificación de la resistencia hidráulica y se determina en función de los factores que determinan la rugosidad del cauce. En el trabajo se diferencia entre el coeficiente de Manning en el cauce principal y en la planicie, donde el nivel de resistencia en la dirección de escurrimiento es mucho mayor. Los albardones son formaciones naturales generadas por los sólidos depositados en las márgenes del cauce principal durante las crecidas. Ambos parámetros son considerados críticos en la simulación y contienen un considerable grado de incertidumbre en sus valores. Esta decisión se toma siguiendo las recomendaciones de los ingenieros del INA.

La salida que provee el simulador con el modelo del Paraná es un conjunto de archivos con los datos simulados (hidrogramas de alturas y caudales) para 15 estaciones de seguimiento ubicadas a lo largo del cauce. La Figura 1 muestra un esquema de la organización de la entrada y salida al simulador Ezeiza.

3 Metodología

El trabajo que se presenta está enfocado en tratar el problema de la incertidumbre de los parámetros de entrada mediante la implementación de un método

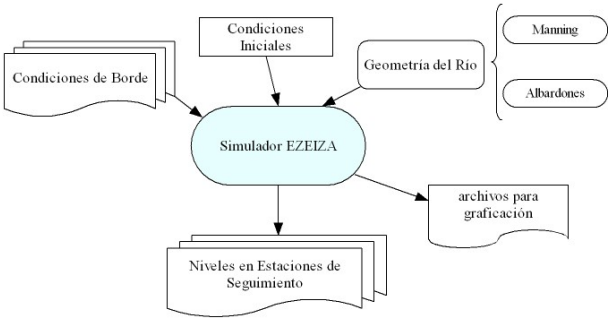


Fig. 1. Esquema de Entradas y Salidas al Simulador Ezeiza

de ajuste de parámetros. Se busca encontrar el mejor escenario, o sea el que minimiza las diferencias entre datos reales (datos observados), y los resultados del simulador (datos simulados). La idea principal del proceso de sintonización aplicado al simulador de cuencas de ríos, se puede ver en la Figura 2.

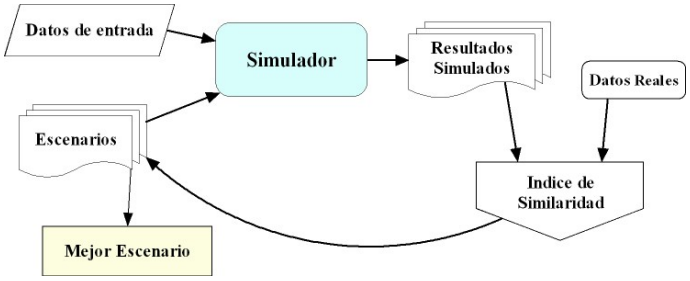


Fig. 2. Esquema del Proceso de Sintonización

La fase de ajuste se hace mediante la implementación de una experimentación paramétrica. Este proceso consiste en lanzar tantas simulaciones como combinaciones posibles de los valores de los parámetros que utiliza el simulador, siendo el objetivo alcanzar el mejor escenario.

En nuestro modelo, el dominio del río Paraná está dividido en 76 secciones y cada una de ella está divida en subsecciones. Se puede encontrar entre 3 y 11 subsecciones, según el ancho de la planicie de inundación en cada sección. El coeficiente de Manning se considera en cada sección, y en cada una de sus subsecciones, en las cuales es tratado como si fuese un parámetro diferente a los efectos de realizar la simulación paramétrica.

Este proceso requiere establecer un *índice de similaridad*, el cual constituye una métrica para medir la diferencia entre los datos reales y los simulados. La

simulación paramétrica permitirá calcular este índice para cada escenario y encontrar el mejor índice de similaridad entre el simulador y el sistema real. La experimentación se realiza con los datos provistos por el INA. Estos son:

- Los datos reales: las alturas diarias del río Paraná en diferentes períodos, que van desde el año 1994 al 2011, y en cada estación de seguimiento a lo largo del cauce considerado.
- Los datos del modelo: condiciones iniciales en todos los puntos del dominio, series temporales de alturas y caudales correspondientes a las condiciones de borde, la información sobre los parámetros en cada sección y los datos de la geometría del sistema real. Estos datos fueron recibidos del INA, ya mejorados, luego del estudio de rendimiento del 2011 [6].

3.1 Generación de Escenarios

El simulador Ezeiza permite modificar los valores de los coeficientes de Manning en todas las secciones y subsecciones en que se divide cada sección, y las alturas de los albardones en sus archivos de entrada, haciendo muy fácil llevar adelante la experimentación paramétrica. La cantidad de escenarios posible está determinada por la cardinalidad de cada uno de los N parámetros considerados, ésta se denomina C_i, donde *i* identifica cada uno de los parámetros. Cada parámetro debe tener asociado un rango de valores propio de su dominio más el paso del barrido con el que se recorrerá dicho intervalo ($Incremento_i$). Para cada parámetro i se puede representar su intervalo e incremento asociados como la tupla:

$$< [Cota_{inf}, Cota_{sup}], Incremento_i >$$

A continuación, se muestra la relación entre estos parámetros, propios de la experimentación paramétrica [3]. El intervalo de barrido está acotado por $Cota_{inf}$ y $Cota_{sup}$

$$\begin{aligned} &\#Escenarios = \prod_{i=1}^{N} C_i donde \\ &C_i = ((Cota_{sup} - Cota_{inf}) + Incremento_i)/Incremento_i \end{aligned} \tag{1}$$

En este modelo se cuenta con 76 secciones con sus valores de Manning. En cada sección se combinan los valores de Manning para planicie y para cauce. Se verá después que en planicie, Manning puede tomar uno de 6 valores diferentes y en cauce, uno de 9 valores. Al cálculo debemos agregar 57, de las 76 secciones que registran su valor de albardones. De este análisis resultan 76 x 6 x 9 x 57 = 233928 escenarios.

La medida tomada como índice de similaridad es el error relativo de los valores simulados respecto a los valores reales. Se obtiene un índice por cada combinación de parámetros, el que tenga el menor valor se corresponde con el mejor escenario de todos. Por otro lado, se calcula el índice de similaridad resultante de correr la simulación con el escenario utilizado por los expertos del INA, con el objetivo de compararlo con el de esta experimentación. Esta fase de ajuste de parámetros se hizo con tres estaciones de prueba, seleccionadas de las 15 estaciones de seguimiento del río Paraná. Para cada una se seleccionó el mejor escenario [9].

Los valores simulados y reales que se comparan son las alturas del río en las estaciones consideradas. Se denonina A_R^k y A_S^k a la altura real y simulada, respectivamente, correspondientes a la *Estación k*. El índice de similaridad, $Indice_j^k$, resultante para la *Estación j* luego de correr la simulación para el *Escenario k*, surge de la siguiente ecuación:

$$Indice_j^k = \left| A_R^k - A_S^k \right| / A_R^k \tag{2}$$

El mejor escenario, que se denomina como $\widehat{Esc_j}$, es el que obtiene el mínimo índice de similaridad. Se calcula como el mínimo índice de todos los escenarios, medido en la Estación j, o sea:

$$\widehat{Esc} = Min_k(Indice_j^k) \tag{3}$$

3.2 Experimentación

En este apartado se presenta un estudio acotado del modelo y su implementación en Ezeiza, con la finalidad de presentar la técnica computacional de sintonización del simulador. La idea es mostrar que se logra el objetivo en tres estaciones de seguimiento demostrando que el método de sintonización es factible y extensible a todo el espacio de parámetros e intervalos.

La cantidad de escenarios se limita tomando sólo los valores del coeficiente de Manning como parámetro crítico, en esta etapa no se consideran los albardones. También se acotó la cantidad de Secciones, tomando una muestra representativa de 3 de las 76 posibles, ubicadas en el área de Paraná Alto (Sección 76), Medio (Sección 36) y Bajo (Sección 01). Sobre estas Secciones se efectuó la experimentación paramétrica con los escenarios que se ven en la Tabla 2.

La cantidad de escenarios se puede calcular, utilizando la Eq. 1, de la siguiente manera:

$$\#Escenarios = 3.2.4.1 = 24 \tag{4}$$

– Cardinalidad Manning Planicie: ((0.2 - 0.1) + (0.1) / 0.1 = 2

Tabla 2. Valores de los parámetros de Manning en cada escenario considerado para la simulación paramétrica.

Número de *Escenario*	*Manning Planicie*	*Manning Cauce*	Número de *Escenario*	*Manning Planicie*	*Manning Cauce*
1	0.1	0.01	13	0.2	0.01
2	0.2	0.01	14	0.2	0.02
3	0.1	0.02	15	0.2	0.03
4	0.2	0.02	16	0.2	0.04
5	0.1	0.03	17	0.1	0.01
6	0.2	0.03	18	0.1	0.02
7	0.1	0.04	19	0.1	0.03
8	0.2	0.04	20	0.1	0.04
9	0.1	0.01	21	0.2	0.01
10	0.1	0.02	22	0.2	0.02
11	0.1	0.03	23	0.2	0.03
12	0.1	0.04	24	0.2	0.04

- Cardinalidad Manning Cauce: ((0.04 – 0.01) + 0.01) / 0.01 = 4
- Cardinalidad Secciones: ((3 – 1) + 1) / 1 = 3
- Cardinalidad Albardones: 1

En la etapa de ajuste de parámetros, se implementó el programa Ezeiza con los datos de cada una de las 24 configuraciones correspondientes a los escenarios descritos. Para medir los errores de predicción se seleccionaron 3 estaciones (de las 15 de seguimiento): Hernandarias (Est1), Bella Vista (Est2) y Rosario (Est3). Esta simulación se realizó con 2000 pasos de tiempo (2000 días a partir del 01/08/1994) y se obtuvieron mejoras significativas en la predicción en los escenarios que se muestran en la Tabla 3, los cuales se seleccionaron tomando el menor índice de similaridad alcanzado. Los coeficientes de Manning que se muestran corresponden al escenario con mejor índice de similaridad para cada estación considerada. En cada caso se muestran dos índices de similaridad cuya explicación es:

- SimulPar: Indice de Similaridad que presenta el mejor Escenario, en la simulación paramétrica, con la Realidad.
- EscINA: Indice de Similaridad del Escenario utilizado por el INA, con la Realidad.

La Figura 3 representa los datos de la Tabla 3 y visualiza la ganancia obtenida con la elección de los escenarios que proveen la mejor sintonización. De los datos de esta tabla, y del gráfico, se puede afirmar que cada una de las

Tabla 3. Escenarios que mejor se ajustan a la predicción en las estaciones seleccionadas.

Estación Seguimiento	$Mejor\ Escenario$	$Manning$ $Planicie$	$Cauce$	$IndiceSimilaridad$ $SimulPar$	$EscINA$
Est1-Hernandarias	5	0.1	0.03	0.0007	0.0526
Est2-Bella Vista	9	0.1	0.01	0.0028	0.0166
Est3-Rosario	18	0.1	0.02	0.0019	0.0684

estaciones consideradas logró mejorar la certeza de sus datos de salida con errores que están por debajo del 0.3%, en cambio en estas estaciones el INA logra su mejor simulación con errores entre 2-7%. Se superó ampliamente la mejora lograda por el INA, luego de su estudio de rendimiento y ajuste de parámetros.

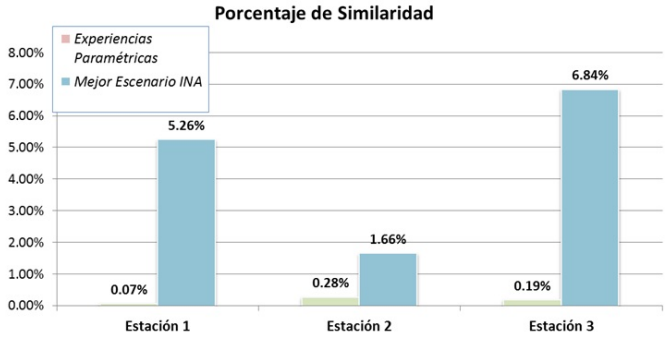


Fig. 3. Comparación entre los porcentajes de similaridad logrados en las simulaciones paramétricas y el mejor escenario utilizado por el INA

4 Resultados de la Experimentación

Este proceso de sintonización se puede extender a todas las otras estaciones a medida que se amplíe la experimentación paramétrica a todos los escenarios-posibles. Esto sucederá al combinar los parámetros de Manning de cauce y planicie de las 76 estaciones en conjunto, y las alturas de los albardones en cada una. Será necesario crear un índice de similaridad que pueda medir los resultados en las 15 estaciones en conjunto. Para llevar adelante esta experimentación se requerirá correr en paralelo todo el proceso. Una simulación completa, para

6000 pasos de tiempo, en un procesador Intel(R)-Dual Core de 1.3GHz, y para un juego de parámetros tarda entre 9 y 10 min. Según vimos anteriormente, tenemos 233.928 posibles escenarios, lo que lleva a más de 2.000.000 de minutos de cómputo para hallar el conjunto óptimo de parámetros mediante una búsqueda exhaustiva. Igualmente son tiempos impracticables y las experiencias futuras de este trabajo, necesitarán llegar al conjunto óptimo, o mediante una heurística aproximarse todo lo posible. Esto es una demostración de que se está ante un problema propio del HPC. Por otro lado, el proceso de sintonización se utilizará para intentar mejorar la predicción del simulador en el futuro, por lo cual será necesario efectuar un proceso iterativo que repita la sintonización de los parámetros, aumentando aun más el tiempo de ejecución del proceso.

5 Conclusiones y Trabajo Futuro

En este artículo se ha descrito una metodología que se aplicará para mejorar la predicción del simulador Ezeiza. Se utilizó un conjunto de experiencias acotadas para probar la factibilidad del método, el cuál proporcionó muy buenos resultados en las tres estaciones que fueron estudiadas. La batería de experiencias que se puso en práctica permitió tener una idea de la necesidad de recursos de cómputo para la etapa siguiente y mostrar que este problema deberá ser resuelto con HPC. Actualmente, se desarrolla la etapa siguiente mediante el uso de procesamiento paralelo. Se implementa un esquema paralelo Master-Worker con asignación dinámica de cargas, utilizando bibliotecas de pasaje de mensajes. El reparto de cargas se refiere a la distribución de sucesivos escenarios a cada nodo worker hasta que finalice el proceso encontrando el mejor escenario. Esta etapa requerirá diseñar una heurística inteligente para continuar con los objetivos del trabajo.

Teniendo en cuenta la importancia de dar alertas hidrológicas sobre crecidas en las cuencas de ríos, el brindar una metodología destinada a mejorar la certeza de los pronósticos de los programas de simulación sería un aporte muy valioso para los expertos que utilizan diariamente estas herramientas computacionales.

Referencias

1. Abdalhaq B., Cortés A., Margalef T., Luque E.: Accelerating Wildland Fire Prediction on Cluster Systems. International Conference on Computational Science (2004) 220–227
2. Balica S. F., Popescu I., Beevers L., Wright N. G. Parametric and physically based modelling techniques for flood risk and vulnerability assessment: A comparison, Environmental Modelling & Software. Elsevier Science Publishers. **41** (2013) 84–02.
3. Bianchini G., Cortés A., Margalef T., Luque E. S2F2M – Sistema Estadístico para la Predicción de Incendios Forestales. I Congreso Español de Informática. CEDI 2005. Granada (España). ISBN: 84-9732-430-7 (2005) 623–629.

4. Jaime, P., Menéndez, A. Modelo hidrodinámico del Río Paraná desde Yacyretá hasta la ciudad de Paraná. Instituto Nacional del Agua. Secretaría de Recursos Naturales y Desarrollo Sustentable. LHA01-165-97. (1997)
5. Krause, P., Boyle, D. P., Bäse F. Comparison of different efficiency criteria for hydrological model assessment. European Geosciences Union. **5** (2005) 89—97.
6. Latessa G. Modelo Hidrodinámico del Río Paraná para Pronóstico Hidrológico: Evaluación del Performance y una Propuesta de Redefinición Geométrica. Tesis de Grado. Fac. de Ingeniería. UBA-INA. 2011.
7. Menéndez A. EZEIZA V: Un programa computacional para redes de canales, Mecánica Computacional. Instituto Nacional de Ciencias y Técnicas Hídricas. **16** (1996) 63–72.
8. Pappenberger F., Beven K., Horritt M., Blazkova S. Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. Journal of Hydrology. **302** (2005) 46–69.
9. Taboada M., Cabrera E., Luque E. A Decision Support System for Hospital Emergency Departments built using Agent-Based Techniques. Practical Application of Agents and Multiagents Systems. ISBN: 978-3-642-19874-8. PAAMS 2011. (2011)

Multithreading model for evacuations simulation in emergency situations

Pablo CristianTissera[1], A. Marcela Printista[1,2], Emilio Luque[3]

[1] Departamento de Informática, Laboratorio de Investigación y Desarrollo en Inteligencia Computacional. UNSL.
[2] Conicet - CCT San Luis.
[3] Departamento de Arquitectura y Sistemas Operativos
Universidad Autónoma de Barcelona - España
{ptissera, mprinti}@unsl.edu.ar
emilio.luqe@uab.es

Abstract.
Evacuation simulations allow to consider preventive measures against possible emergency scenarios. We have developed a simulation model that takes into account not only the characteristics of the environments but also is able to represent social behaviours that would render our models more accurate and realistic. The proposed model has a hybrid structure where the dynamics of fire and smoke propagation are modelled by mean of Cellular Automata and for simulating people's behaviour we use Intelligent Agents. In this paper, a behaviour in panic situation is added to the existing ones. Moreover, as main contribution, this paper explains the implementation of the model in which we apply a functional decomposition in order to accelerate the simulation and take advantage of current computer architectures.

Keywords: Evacuation Simulation, Social Behaviours, Cellular Automata, Intelligent Agents. Multithreading.

1 Introduction

In the last years, several modelling approaches have been proposed to deal with the emergency evacuations because the prediction of the people's behaviour is of great public interest. Models used for evaluating the evacuation processes can broadly be categorised in microscopic and macroscopic approaches. The macroscopic approaches are based on differential equations that take into account the similarities with systems previously studied like dynamics of fluids. On the other hand, the microscopic approaches allow to investigate how the system state evolves during the model runs. References to different models may be obtained from [2].

We developed a hybrid model where the environmental dynamics are modelled by means of Cellular Automata (CA), because it are suitable for modelling process of diffusion like fire and smoke and for simulating people's behaviour we are using the intelligent agent (IA) concept. We used a behaviour-based agent architecture because it allows us to work with behaviours beyond than those purely reactive[10,11]. This

type of system provides solutions in dynamic and uncertain environments, where the agent has only a partial view of the problem.

The proposed simulation system allows to specify different scenes with a large number of people and environmental features, making easier the study of the complex behaviours that arise when the people interact. Our proposal could be used by architects, government agencies, foundations, etc. in order to know the security threats of a possible disaster, help with appropriate actions of prevention for a quick and efficient way to evacuate a building through the design of active policies that minimize the evacuation time when circumstances require it.

Our model is a process that consumes a significant amount of time to simulate a complete evacuation when the environment size and / or the number of people is considerable. The emergence of multi core processors introduces a real challenge for parallel applications, that is to exploit such architectures at their maximum potential. This leads us to develop a model to achieve a competitive performance.

Section 2 describes the Hybrid Model for the Evacuation Simulation, by explaining briefly the environmental and the pedestrian sub models. Section 3 explains the implementation of three behaviours commonly observed in emergency evacuation. In Sub-section 3.1 we explain how to perform the association of an agent with a specific behaviour during the execution of the model. In section 4 we present the Multithreading Model for Evacuation simulation. In section 5 we describe our work with different instances of the problem at hand and report the performance analysis of each case and in section 6 discuss the conclusions and future works.

2 Simulation Model

The model consists of two sub-models, called environmental (EsM) and pedestrian (PsM). This model along with the computational methodology allow us building an artificial environment populated with autonomous agents, which are capable of interacting with each other. The Fig. 1 shows the hybrid model.

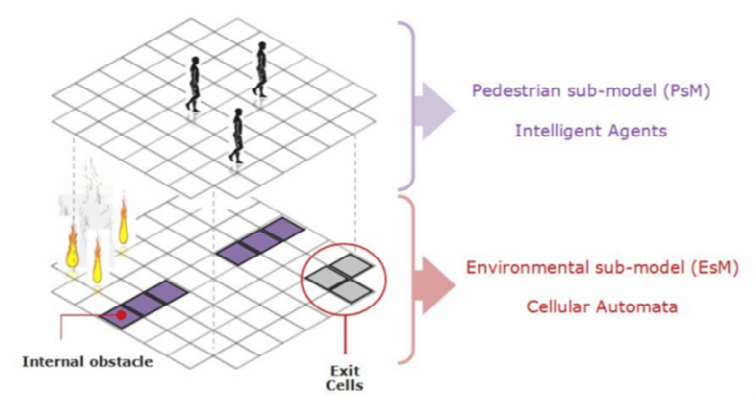


Fig. 1. Hybrid model consisting of environmental and pedestrian sub models

The details of the hybrid model have been reported in [1, 3], and for reasons of space they will not be reproduced in this paper. We briefly mention only those aspects that are necessary to understand the proposal.

The EsM describes the spatial configuration of the environment (geometry of space, exit doors, internal barriers, etc.) and models the processes of diffusion of smoke and fire. The EsM is based on CA, which are discrete dynamic systems that have the capacity to develop complex behaviours from a simple set of rules [9]. Basically, these rules will allow to specify the new state of a cell based on the state of the neighbouring cells.

The PsM uses the concept of intelligent agents to describe the cognitive processes of individual agents and interactions among multiple agents in a specific environment. Through interaction and coordinated evolution of these two sub-models it is possible to obtain a model capable of simulating indoor environments with a finite number of exits that must be evacuated by a group of people due to the threat of fire and the effect of the smoke.

In the proposed model an agent is placed on an environment described by a bi-dimensional grid where they can find different elements such as walls, obstacles, exits, presence of smoke, fire and other agents. The agent architecture is illustrated in Fig. 2 (left).

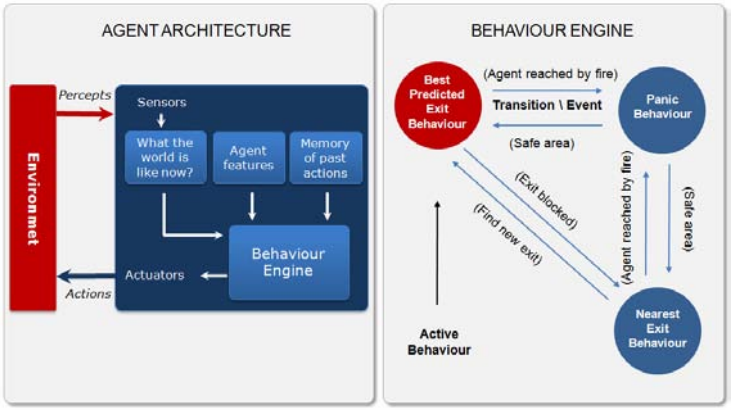


Fig. 2. Agent Architecture and Behaviour Engine

In our model, agents respond to a behaviour-based architecture because one of the major drawbacks of this type of system is that multiple behaviours with different objectives may be attempting to take control of the agent at the same time. To solve this problem, known as the action selection problem [7], it is necessary to develop a mechanism that allows us to select the appropriate behaviour in a given situation. In our model each agent has an associated behaviour engine, shown in Fig. 2 (right) that manages decision-making processes. This engine is a nondeterministic finite automaton, where each node represents the implementation of behaviour while the transitions represent the event for which the agent can change the state.

This arbitration state-based mechanism [8], selects an appropriate behaviour to deal with the current situation from a determinate event detected in the environment

[7]. In this way, an agent can change its behaviour during execution of the model according to a predetermined set of rules that serve as triggers for this change.

3 Primitive behaviours

In the current state of development, the simulator has the capability to implement three behavioural categories.

In the *Nearest Exit behaviour* (NE_B), the agent will try to get out the exit closest to its current position. In this behaviour the decision process will take into account the position of the agent, the direction toward the nearest exit, the state of its environment in relation to the progress of fire and smoke, but it ignores information from other alternative solutions, the behaviour of other agents and it will not take unexpected or altruistic decisions.

In the *Best Predicted Exit behaviour* (BPE_B), the agent will analyse different exits and choose one that it predicts the fastest exit to evacuate. The decision process will take into account the position of the agent, the state of its environment in relation to the progress of fire and smoke, the distance to alternative exits, the density of crowd trying to evacuate for each exit (only if the agent can see the exit) and the stress level in relation to its tolerance to it. As the evacuation progresses, the agent is predicting the cost (in time) to evacuate by each of the exits that are available in the environment. The inferred lower cost will indicate the best exit. For that, the decision-making process evaluates a cost function that indicates which is best exit.

The resulting procedure instructs the agent to which exit to go. In addition, the procedure involves two dynamic factor used to adjust the number of times the agent executes the action to evaluate the exits[1]. This is done to limit the effect of indecision of the agents. This factor depends of a environment size and it is dynamically adjusted according the time elapsed since the start of the evacuation.

Finally, in the *Panic behaviour* (P_B), unlike the previous behaviours, the agent does not realize any type of analysis over the exits. An agent assumes this behaviour in a situation of extreme danger, when the agent's current position has been achieved by the spread of fire. In such situation, the agent only analyzes their position and the state of their environment in relation to the progress of the fire and smoke. The decision-making process of the agent evaluates the condition of the cells in its proximity searching cells that remove it from the fire, without mattering if these cells offer it or not a better position respect of the exit chosen, that is to say, the agent tries to escape and to reach a sure position without presence of the danger.

3.1 How does the behaviour engine work?

So far we have only defined the behaviours implemented in the model, but we have said nothing yet about how these are related through the behaviour engine, giving origin to agents who can change its behaviour along the simulation. Before, it should be noted that the agents at the beginning of the simulation have an assigned behaviour, but as the simulation progresses, could arise different situations in which it is suitable or even imperative change the behaviour of the agent. Next, we will discuss the *Events* that can lead to an agent to change their behaviour:

Agent Reached by Fire and Agent in Safety Zone: Any agent reached by the spread of the fire along the simulation (agent reached by fire event), regardless its behaviour, detects the situation and changes its current behaviour by the P_B with the purpose of going out of the situation of danger of in an immediate way. Once the agent has reached a safe area, the same resumes its original behaviour (safety zone event) with the objective of continuing the evacuation by the selected exit.

Blocked Exit: An agent whose behaviour is BPE_B, uses two parameters to determine the amount of inferences that the agent can perform and how often these inferences can be done [1]. It is possible that along the evolution of the model, an agent has made all possible inferences and taken its last decision, therefore, the exit to which it is addressed cannot change, but it can happen that after taken the final decision, the exit will be blocked due to the spread of the fire, then the agent already cannot evacuate for this exit. When this situation is detected an agent changes its behaviour by NE_B, because it must select a new exit to evacuate but has already exhausted all the instances that allowed it choose an exit.

Select New Exit: An agent whose behaviour is NE_B, along the evolution of the model may be in a situation where it cannot move towards the selected exit because this exit is too congested. When the agent detects this situation and according to its stress level may or may not change their behaviour in a probabilistic way by BPE_B, with the purpose of find a new exit that allow it evacuate more quickly.

Although, currently in our implementation only we have defined a few events of change of behaviour, it is necessary to emphasize that our model allows easily add new events and new behaviours that will allow describe better the reality and therefore improve the model in a progressive way.

4 The Multithreading Model

The model proposes the execution of so many time steps as necessary until the last alive individual in the environment has been evacuated. The model evolves to discrete steps of time, which leads to discretize the progress of an individual pedestrian, however the movement of a crowd should appear as a continuous phenomenon.

To solve the problem caused by the discretization, our model introduces the execution of sub-time steps between two consecutive time steps [2,3]. As can be seen in Fig. 3, in every sub time step, five phases are executed (*Environmental phase, Phase of Intentions, Phase of conflicts resolutions, Phase of propagation of responses and Phase of updating of the agents)*. In the following, we give a short description of each phase.

Environmental phase: Is responsible to evolve the environmental sub model. In this phase, the evolution rules of the CA for the spread of fire and smoke are applied. After that, this phase should also re calculate the distances from each cell to each exit, due to the spread of fire modifies the environment in which agents must find the way to the exits.

Phase of intentions: this is the first phase in the pedestrian sub model evolution. During this phase, each agent writes a intention of movement in the cell to which one wants to move (target cell). The decision of which is the target cell is determined by

current behaviour of the agent. It should be noted that a target cell can be empty or occupied by another agent and can also receive more than one intention since more than one agent may intend to move to it.

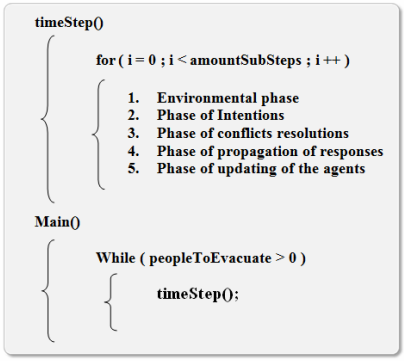


Fig. 3. Main structure of the model.

Phase of conflicts resolutions: This stage is responsible for resolving existing conflicts in the cells with more than one intention of movement. During this phase the agents will receive a first response to its request for movement. The response that each agent will obtain can be {*accepted movement, denied movement, uncertain movement*}.

If cell in conflict is empty: the conflict resolution process selects an agent between all candidates to occupy the cell in conflict in the next sub time step[1], therefore this agent will obtain a *accepted movement* response, while the rest of the candidates will obtain a *denied movement* response. The process gives priority to the selection of agents with greater speed and fewer points of damage (agent parameter). If the conflict persists, the selection will be random.

If cell in conflict is currently occupied by another agent: the process will check if the cell could be free in the next sub time step. If it will be free, the same procedure of the empty cell is executed. If there is no possibility of the cell to be unoccupied, all agents candidates will receive the response of *movement denied.*

Finally, in the case that it is no possible determine if the target cell will be free in the next sub time step, due to the fact that the movement of the agent depends on the response of another cell currently occupied by another agent and so on, the agent receives as response to its request *uncertain movement*. As the simulation progresses, the possibility of occurrence of this case increases, since the agents tend to gather in crown in the vicinity of the exit and therefore their movements depends on people who are several positions later. This type of conflict is solved by the following phase.

Phase of propagation of responses: The responsibility of this phase is the propagation of received responses by agents in the previous phase. In this way if an agent has received as response *uncertain movement*, at this stage its movement is accepted or denied.

During subsequent sub steps, all agents with *uncertain movement* will remain in this state. Once one of them receives an *accepted* or *denied movemen*t in some sub step, then it will start a backward propagation of novelty, resolving several conflicts in the process. With the purpose to make it clear the operation of this phase we will exemplify different situations that can occur. Suppose that we will call A to an agent that attempting to move into a cell occupied by an agent B, which in turn wants to move to a cell occupied by an agent C. Clearly the agent A cannot move because it is not possible to determine if the agent B will move. In a similar way the same thing happens with the agent B. But once the agent C receives its response of movement accepted or denied, this will spread its response to agent B which can propagate its response to the agent A. In this way the conflict can be solved. Now, suppose the same previous situation, but with the difference that the agent C wants to move to the position of the agent A. In this situation we are in the presence of a cycle and therefore a deadlock situation since the agents will not move because they are waiting for a response that will never come. To solve this deadlock situation, this phase can detect the cycles and all the agents in a cycle receive a *denied movement* response. At first sight the answer given to the agents may seem arbitrary, why the agents did not get a *accepted movement* response?. This is so, because there can be agents who try to move to a position occupied by other agent which is in a cycle, but these agents do not belong to the cycle. This situation can have a large number of variants when working with thousands of agents, therefore it seems reasonable to give *an movement denied* response, since in the next sub times steps the conflicts will be solved.

Phase of updating of the agents: Finally once all agents have its answer, the position of the agents is updated.

It is important to emphasize, that there is a clear division of tasks between the phases mentioned above. While the environmental phase is responsible to carry out the evolution of the environmental sub model, the four remaining phases are responsible for evolving the pedestrian sub model.

Our proposal is aimed at carrying out a parallel shared memory model where it is possible to perform task-level parallelism, since a set of threads con solve the environmental sub model while another set of threads solve the pedestrian sub model. The multiple threads assigned of the same task assist in the resolution of disjoint areas of the grid in a data parallelism way.

To do this then we will see how to perform the update of the cells in each sub time step.

As mentioned above, our model uses a CA (sub environmental model), where agents are positioned. At the time to evolve a CA, it is necessary to have an auxiliary structure which saved the next states of all cells of the automaton solved by means of the application of the rules of evolution. In this way, will we have then two CA, the first will represent the state at time T of the CA, while the second (in built) represents the condition of the CA at time $T+1$. Once completed the process, the new representation becomes the new current state of the automaton and the process repeats. In the case of the sequential implementation of our model, there are no problems at the moment of updating of the cells. This is because when the pedestrian sub model begins with the execution of its phases, the environmental sub model has been fully resolved. It is important to emphasize this point, because rarely it is expected that an individual try to perform a move to a position occupied by the fire.

This is achieved because the environmental model has been solved and therefore it is possible for the agent to be able to determine its next position by looking at the state of the automaton at the next time step to avoid cells with fire. This present a problem in the Multithreading Model since, both spread of fire as the agent evolution are been executed in parallel.

To solve this situation, our proposal is to advance in a time step the resolution of the environmental sub model. That is, while the threads of the pedestrian sub model use the structure of the time *T* to obtain the pedestrian configuration of the time *T+1*, the threads entrusted to solve the environmental sub model will be using the structure of the time *T+1* to obtain the environmental configuration of the time *T+2*. In this way, is possible that the agents can have a forward vision of the environment already resolved as the case of the sequential implementation of the model, as can be seen in Fig. 5.

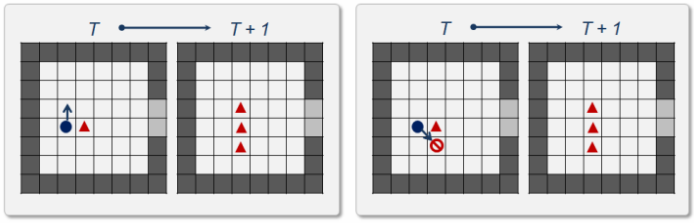


Fig. 4. The agent (circle) in T sees the fire (triangle) in T+1 and then it rules out those cells as next move.

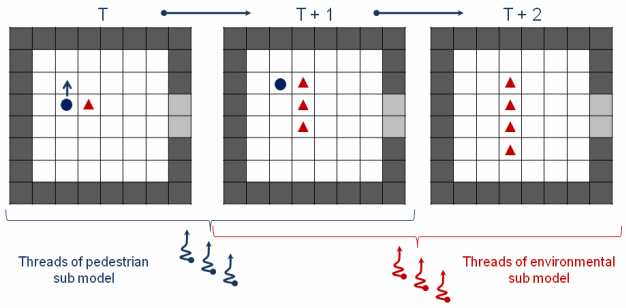


Fig. 5. Threads of pedestrian and environmental sub models.

5 Test-Case Scenario and Results

In this section, we present the simulation results of the explained research. The experiments were carried out with EVACOMP, a hybrid simulation system based on cellular automata and intelligent agent. EVACOMP is a system developed in C and OpenMP and uses the graphical interface (off line) of EVAC Simulator [1,2]. The experiments consider two environment configurations of the buildings to be evacuated (*A, B*):

- *A, 60 × 30 m^2*, three exits of 4 m. each and 2500 pedestrians distributed evenly (50% NE_B and 50% BPE_B).
- *B, 80 × 60 m^2*, three exits of 6 m. each and 4500 pedestrians distributed evenly (50% NE_B and 50% BPE_B).

The experiments are designed to test the performance of the EVACOMP vs. the sequential implementation of the model. With the purpose of obtaining acceptable statistical data, the results shown below correspond to the average of 50 independent replications of each experiment. All execution times values are expressed in seconds and we always set up a thread by core. From the results showed in Table 1, for the Experiment A is possible to visualize, that the best execution times were obtained using 4 threads, and therefore it is where the major speedup is obtained. In this case, while it is possible to see a reduction in execution times, the speedup obtained seems relatively mild. We develop the second experiment, where we increase the size of the environment and the quantity of individuals with the purpose of increasing the quantity of work necessary to solve the model.

Table 1. Execution Times, Speeup and Efficiency values for experiments A and B.

	Sequential	***Threads 2***	***Threads 4***	***Threads 8***	***Threads 16***
Execution Time Experiment A	27,37	19,29	17,16	17,75	22,74
Execution Time Experiment B	439,85	268,12	170,55	134,19	206,41
Speedup Experiment A	X	1,41	1,59	1,54	1,20
Speedup Experiment B	X	1,64	2,57	3,27	2,13
Efficiency Experiment A	X	70,91 %	39,86 %	19,26 %	7,52 %
Efficiency Experiment B	X	82,02 %	64,47 %	40,97 %	13,31 %

As we see in Table 1, the execution times for Experiment B increased significantly compared to the first experiment. Here, the best execution times were obtained using 8 threads, and therefore it is where the major speedup is obtained. It is important to highlight for this case, that the obtained speedup is acceptable and in addition it is better than the speedup obtained for the best case of the first experiment. By making this comparison, it is possible to appreciate, that both the speedup and the efficiency obtained improve for the second experiment, which is a good indication of which on having increased the load of work in the system the performance of the parallel model improves. Because to the orientation of this work and for reasons of space, we do not report the information about times of evacuation and travelled distances by the individuals. A wide series of experiments can be consulted in [1,2,3,4], where the empirical values obtained for the evacuation time are comparable to other implementations, which have validated their results against real evacuation exercises [5] [6]. For the experiments, we used a multicore equipment with 4 processors AMD

Opteron 6128, 2.0GHz (8C), and RAM memory of 64GB Memory (16x4GB), 1333MHz.

6 Conclusions and Future Works

We presented a model capable of simulating indoor environments with a finite number of exits that must be evacuated by a group of people due to the threat of fire and the effect of the smoke. The proposed model consists of two sub-models, the Environmental Model (EsM) and Pedestrian Model (PsM). The EsM, based on CA, manages the spatial configuration of the environment and models the processes of diffusion of smoke and fire. The PsM is the part of the hybrid model focuses on representing the human behaviours.

The proposed model to perform both task-level and data-level parallelism, where a group of threads will be responsible to evolve the pedestrian sub model pedestrian and another group of threads will be responsible to evolve the environmental sub model. While our development is not yet complete, the results of our Multi Threading implementation presented here are encouraging.

As future works, it is important to decide the optimal size of each set of threads to solve each sub model. Our model is going to use some strategy that will enable us to achieve an optimal balance in the allocation of threads to the resolution of each sub model.

References

1. P. C. Tissera, A. Castro, A. M. Printista, E. Luque, Evacuation Simulation Supporting High Level Behaviour-Based Agents, Procedia Computer Science Vol. 18, 2013, pp. 1495-1504, proceedings of the International Conference on Computational Science, ICCS 2013.
2. P. C. Tissera, A. M. Printista, E. Luque, A hybrid simulation model to test behaviour designs in an emergency evacuation, Procedia Computer Science Vol. 9, 2012, 266-275, proceedings of the International Conference on Computational Science, ICCS 2012.
3. P. C. Tissera, A. M. Printista, E. Luque, Implementing sub steps in a parallel automata cellular model, in: Computer Science and Technology Series-XVII Argentine Congress of Computer Science-Selected Paper, 2012, pp. 81–93.
4. P. C. Tissera, A. M. Printista, M. Errecalde, Multi-column partitioning for agent-based ca model, in: HPC Proceeding. JAIIO, SADIO-Argentina, 2011.
5. Aik Lim Eng. Exit-selection behaviors during a classroom evacuation. International Journal of the Physical Sciences, Vol. 6 (13), 2011, pp. 3218–3231.
6. Klupfel Hubert Ludwig. A Cellular Automaton Model for Crowd Movement and Egress Simulation. PhD thesis, Universitat Duisburg-Essen, 2003.
7. P. Pirjanian, Behavior coordination mechanisms - state of the art, Tech. rep., USC Robotics Research Laboratory, University of Southern California, 1999.
8. A. Saffotti, The uses of fuzzy logic in autonomous robot navigation, Soft Computing Vol. 1 (4), 1997, pp. 180–197.
9. Wolfram Stephen, Cellular Automata and Complexity, Addison Wesley, USA, 1994.
10. P. Maes, The dynamics of action selection, in: IJCAI-89, MI, 1989, pp. 991–997.
11. R. A. Brooks, A robust layered control system for a mobile robot, IEEE Journal of Robotics and Automation, 1986, pp. 14–23.

Efficiency analysis of a physical problem: Different parallel computational approaches for a dynamical integrator evolution.

Adriana Gaudiani[1], Alejandro Soba[2,3], and M. Florencia Carusela[1,3]

[1] Instituto de Ciencias, Universidad Nacional de General Sarmiento, Los Polvorines, Argentina
agaudi@ungs.edu.ar
[2] Comisión Nacional de Energía Atómica
Buenos Aires, Argentina
[3] Conicet, Argentina

Abstract. A great challenge for scientists is to execute their computational applications efficiently. Nowadays, parallel programming has become a fundamental key to achieve this goal. High-performance computing provides a solution to exploit parallel architectures in order to get optimal performance. Both parallel programming model and the system architecture will maximize the benefits if both together are suitable to the inherent parallelism of the problem.
We compared three parallelized versions of our algorithm when applied to the study of the heat transport phenomenon in a low dimensional system. We qualitatively analyze the obtained performance data based on the own characteristics of multicore architecture, shared memory and NVIDIA graphical multiprocesors related to the traditional programing models provided by MPI and OpenMP, and Cuda programming environment.
We conclude that GPUs parallel computing architecture is the most suitable programing model to achieve a better performance of our algorithm. We obtained an improvement of 15X, quite good for a program whose efficiency is strongly degraded by an integration process that essentially must be carried out in a serial way due to the dependence of the data.

1 Introduction

To analyze the dynamic evolution of a real system requires the development of a model that should include among other things the characteristic times of the phenomenon under study. Generally, these models are expressed as differential equations. Depending on the type, these equations are integrated using different numerical integration methods. This requires discretizing the equations, where the length of the integration step δT must be directly related to relevant timescales of the real physical problem, and the total time $t = \delta T.T$ (T total number of integration steps) must correspond to times longer than the typical duration of the phenomenon under study. This is a necessary condition to achieve an adequate insight of the problem. However, t has not necessarily

a direct correlation with the real time of the simulation, which depends on the characteristics of the algorithm implemented.

In particular, parallel algorithms discussed in this paper are applied to study the phenomenon of heat transport in one-dimensional devices immersed in thermal environments. A serial execution of this algorithm is executed in times measured in days. Thus in order to optimize the computational resources and to reduce the large execution times required in the case of a simple serial integration, we propose different parallel implementations. We present schematically the main idea in Figure 1

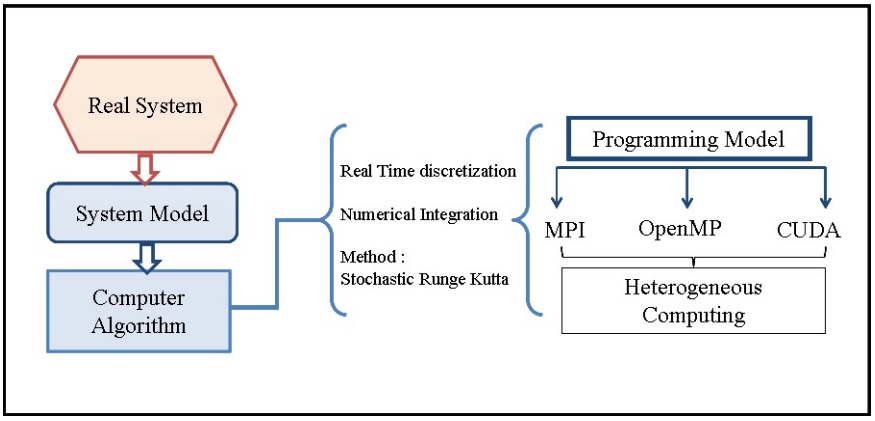


Fig. 1. The major steps from problem to a parallel computational solution.

In this paper, we describe the performance results that we obtained of our parallel algorithm implementation, using three parallel programming models and giving an overview of the heat algorithm behavior in each of our parallel implementations. They are message passing model with MPI on distributed memory systems, shared memory with OpenMP on multicore systems and finally, data parallelism on graphic processing units (GPUs) in a single instruction-multiple thread (SIMT) architecture.

Organization: This paper is organized as follows. In section 2 is given an overview of heat transport algorithm and parallel programming models used to improve the algorithm, along with the desired features on applications that will benefit from these High Performance Computing Systems (HPCS). In section 3 we discuss the performance of our parallel implementations and compare the

programming effort and resulting performance. In section 4 and 5 we present the experimental results and conclusions of our work respectively.

2 Background

In this section we first present the general features of the heat transport algorithm, then we highlight those desired aspects for efficient execution of parallel algorithms as a function of the target architecture.

2.1 The heat transport algorithm

The device is modeled with two chains of $N/2$ atoms with harmonical nearest neighbors interactions with a strengh constant $ki, i = 1, N + 1$. x_i denotes the displacement from the equilibrium position of each particle. The system properties are obtained in the stationary regime, that is when the system thermalize. If the integration is made with the stochastic Runge-Kutta (SRK) [3] algorithm and for the chosen time step (discussed below), the last condition is fulfilled for $T > 10^8$ integration steps [1]. Moreover, we are also interested in the effect that the size system N has on the thermal properties of the physical system. But, as the thermalization time also depends on the number of atoms N, the size study requires to increase the integration time and the memory resources.

The dynamical equations can be written in a non dimensional form as:

$$\begin{aligned} \ddot{x_i} &= F(x_i) + k_i(x_{i+1} + x_{i-1} - 2x_i); \qquad i = 2...N-1 \\ \ddot{x_i} &= F(x_i) + k_i(x_{i+1} + x_{i-1} - 2x_i) - \gamma\dot{x_i} + \sqrt{(}2\gamma K_B T_{L,R})\xi_i(t); \qquad i = L, R \end{aligned} \tag{1}$$

$\dot{x_i}, \ddot{x_i}$ are the velocity and acceleration of the atom i respectively, $F(x_i)$ is a periodic on-site potential that model a substratum and the last term corresponds to the harmonic interaction.

The system is driven out of equilibrium by two mechanisms:

a) The ends of the segments are in contact with Langevin type thermal reservoirs with zero mean and variance $< \xi_{L/R}(t), \xi_{L/R}(t') > = 2\gamma K_B T_{L,R}\delta(t - t')$, where γ is the strengh of the coupling between the system and the baths. The temperatures of the L/R (left/right) thermal baths are simultaneously modulated in time with frequency ω_{temp}: $T_{L,R}(t) = T_{0,i}(1 + \Delta sgn(sin(\omega_{temp}t))$, $i = L, R$, where $T_{0,i}$ is the reference temperature of each reservoir and $T_{0,i}\Delta$ is the amplitude of the modulation. Only the atoms at the ends are inmersed in the thermal reservoirs.

b)The modulation of the coupling between the two segments is given by $k_{N/2}(t) = K_0(1 + sin(\omega_K t))$, with frequency ω_K.

The dynamical evolution is obtained integrating the system given in Eq.1 with a stochastic Runge-Kutta algorithm (SRK), with an integration steps $\delta T = 0.005$. This time step corresponds to a physical time greater than the relaxation time of the reservoirs.

The thermal properties of the system are obtained in the stationary regime, that is when the system thermalize. This condition is fulfilled if the integration with the SRK method and the chosen time step is made for $T > 10^8$ integration steps. More over, we are also interested in the size effect (N) on the thermal properties. However as the thermalization time also depends on the number of oscillators N, this study requires an increase of the integration time and the memory resources.

We tailor the heat transport algorithm for the different parallel computing architectures, taking into account that for every time stage of the numerical method, the data for the x_i element depends on x_{i-1} and x_{i+1} at the same time. These interactions are first neighbors type and they are schematically shown in Figure 2

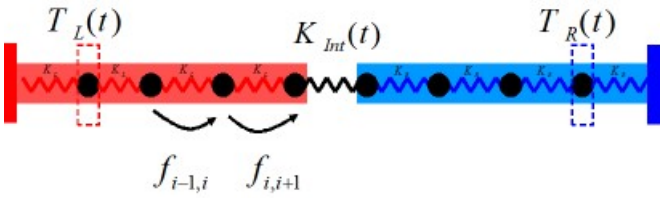


Fig. 2. Schematic diagram of the physical model

For the purpose of study the system behavior we must run several scenarios by combining different values of frequency (ω_K) and temperature (ω_{temp}) parameters. This means investing a longer execution time than for a single pair $(\omega_K, \omega_{temp})$ of parameters. We must consider that the complexity order of the heat algorithm for a single pair is $O(T * 2^N)$, as shown by the author of this paper [4]. Notwithstanding we use chains of N=256 to N=2048 oscillators for this work, we need to increase these values in the physical problem.

2.2 Parallel Programming Overview

Developing a parallel application is strongly conditioned by the system on which this will be deployed and the programming model chosen. But the choice of the model is made in terms of the available parallel computational resources and the type of parallelism inherent in the problem. In parallel computation the most

common alternatives today are message passing, data parallelism and shared memory. In the following, we describe some features of the three standard parallel programming models[2].

MPI is a standard for parallel programming on distributed memory systems, as are clusters. MPI communication libraries provide an interface for writing message passing programs. The most important goals of this model are: achievable performance, portability and network transparency. However, some MPI applications do not scale when the problem size is fixed and the number of core is increased, also they perform poorly when require large shared memory. Our MPI program uses message passing for communications and it was designed in a Single Program Multiple Data (SPMD) way as parallel paradigm. SPMD applications compute a set of tasks and then communicate the results to their neighbors. Just as in our application, tasks need information of their neighbors before proceeding with the next iteration.

The two major hardware trends impacting the parallel programming today are: the rise of many-core CPU architectures and the inclusion of powerful graphics processing units (GPUs) in every desktop computer [6].

OpenMP is a portable approach for parallel programming on shared memory systems that offer a global view of application memory address space, helping to facilitate the development of parallel programs. OpenMP on shared memory systems has become a solution to solve a wide range of large-scale scientific problems which can be solved in parallel to decrease the execution time [8].

GPUs are an inexpensive commodity technology that is already ubiquitous. These technology is highlighted by massively many-core multiprocessors and data level concurrency. It provides a general purpose parallel computing architecture, in the case of modern NVIDIA GPUs it is called Compute Unified Device Architecture (CUDA). In general purpose computing the GPU is used as CPU co-processor in order to accelerate a specific computation. CUDA enables to divide the parallel program execution in tasks that can run across thousands of concurrent threads mapped to hundreds of processor cores. The application benefit with GPUs parallelism when code is written in a Single Instruction Multiple Data (SIMD) mode, this means many elements can be processed in lockstep running the exact same code [9].

3 Parallel Implementations

Our experiences were carried out in a 56 nodes multicluster Intel(R) Dual Core Xeon(TM) 5030 of 2.66GHz processors with a infiniband switch, a multicluster 32 cores Intel Xeon(R) E5-2680 of 2.70GHz and 20MB L2 cache and 64 GB

of RAM memory and a GPU Geforce GTX 560TI Fermi GF114 with capability 2.1 and 1Gb de RAM.

The main computation in the heat transport problem is the SRK algorithm. The numerical method simulates the system evolution over time and it computes each oscillator status at each iteration. This type of computational approach forces a serial integration with no possibility of distributing task between several processes. As we need $T > 10^8$ time steps and oscillator chains with more than 2000 elements, it becomes crucial to minimize the SRK execution time. We address this problem by writing the three versions of our algorithm and then evaluate the performance achieved.

Our first approach was a MPI parallelization when we had to deal with inter-process communication to reduce the overhead. The SPMD implementation works with a simple mapping method by assigning identical amount of data - chain elements - for each process. The features inherent to the problem are the cause of a compulsory data transfer between the neighbours in the chain, because in each integration steps we need data of the last step. This fact is responsable of the bad performance of the MPI implementations (not showed here), where a severe degradation of the speed-up even with two process are obtained, no matters the increment up to N=2000 and despite of the low latency of the network (¡180ns). Expected result, because the kind of algorithms used and to the large number of explicit time steps needed, is neccesary mention that the amount of time needed to perform a complete integration for the whole set is aproximately 0.3 ms.

The next stage was to develop a parallel program for a shared memory architecture with OpenMP. It was a fairly simple task taking advantage that data are read from memory shared by all processes. By avoiding inter-processes message passing we reduced the overhead and got a good performance, as explained in the next section. The pseudocode is shown in Figure 3. The difficulty with OpenMP is that it is often hard to get decent performance, especially at large scale[5].

Finally we present our GPU implementation for the integrator SRK algorithm. We wrote a CUDA kernel to carry out the parallel version taking into account the data dependencies inherent to the problem. The thread tasks are represented by the kernel code and kernels are mapped over a structured grid of threads. The program structure is sketched in Figure 4.

The whole SRK operations performed on each chain element are carried out concurrently by the kernel code launched on every thread. The threads are mapped to processors depending on the grid configuration, that is how many blocks and threads per block are specified when invoking the kernel. It should be noted that the results of the intermediate steps need to be copied to host

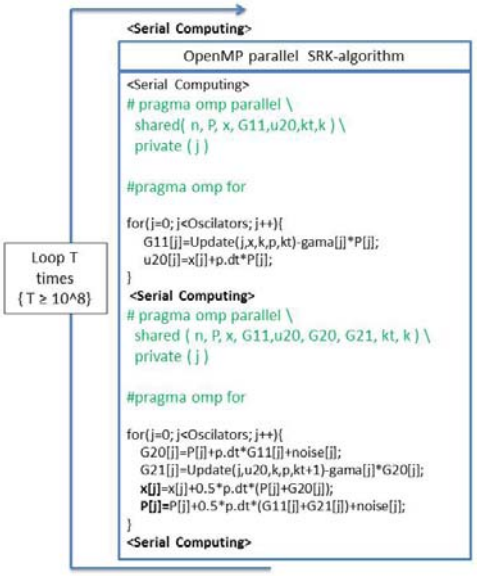


Fig. 3. OpenMP pseudocode program

memory. This copy is performed in ocassion to store partial results, and it is an overhead source. We used Occupancy Calculator spreadsheet to select the best data layout and to maximize Stream Multiprocessors occupancy [7]. In our program, we achieve the best performance when each block has 256 threads.

4 Performance Evaluation

In this section we evaluate the computational performance of heat transport algorithm on the three platforms.

1. MPI parallel algorithm: we analize the system behavior for a number of oscillators between 256 up to 2000. As we mentioned in the earlier section a dramatic performance degradation with the increase in number of processors is visible. Due to the domain division and the requirement in the integration process of interchange the information related to first neighbors in the chain, the amount of communications increases as a result of continually updating the state of those elements. As a consequence of this the performance is not better even in case of add more particles in the chain, because the bottle neck in the code still are the neighbors communications and this number

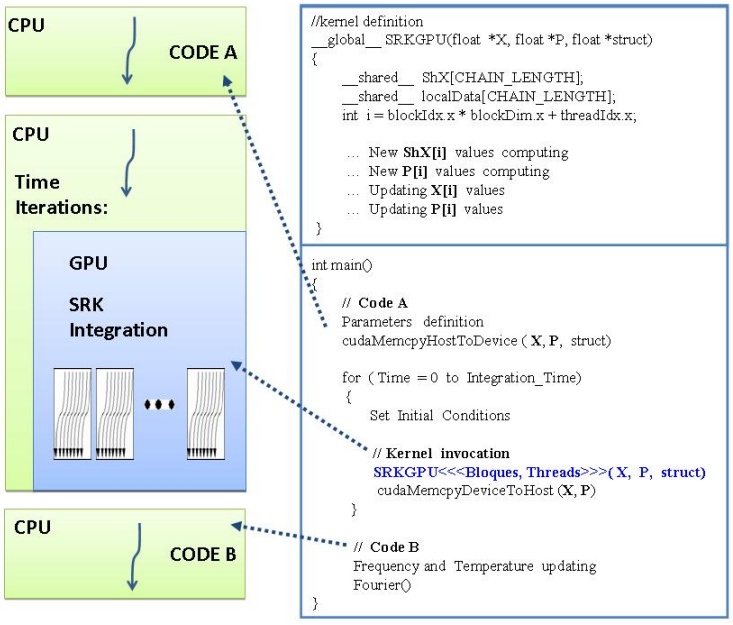


Fig. 4. High Level Structure of CUDA program.

increases with the numbers of domains used. The numerical integrator SRK does not allow the possibility of parallelize the main loop, that is why the performance of this technique is limited to only a few processors. Comparing the estimated latency time (aprox. 180 ns) and the integration of a one single time step of the whole set (aprox. 0.3ms) we are sure than the introduced overhead in the comunication is responsable of this degradation in the speed-up.

2. OpemMp parallel algorithm: We used two multicores cluster, a 32-Core IntelXeon and a 8-Core Xeon. Figure 5 displays the speed-up achieved in both systems for 256 and 1024 chain elements. In this approach, data structures are allocated in shared memory and stay there for every loop during the execution, so we launched a team of threads to parallelize the SRK algorithm. We parallelized those loops in which computation is independently of one another, but still endures the overhead imposed by the serialization of the integration process. The process of updating oscillators states at each iteration was benefited with this programming model. We get a maximum efficiency of 41% for 8 processors and 256 chain oscillators and 66% for 1024 oscillators. The scheduling scheme was delegated to the compiler and runtime system, the results were similar to static and dynamic scheduling.

Figure 5

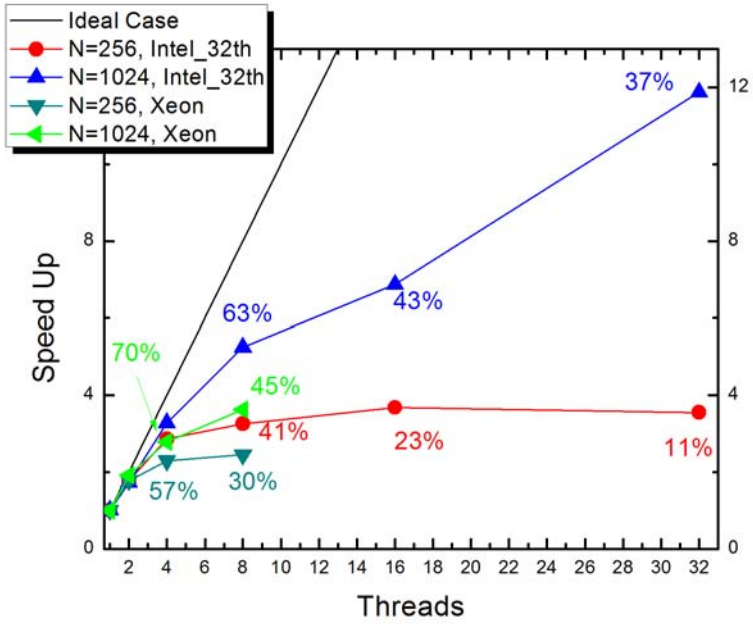


Fig. 5. OpenMP implementation results. The percentages represent the quality of the result with respect to the ideal case i.e. the amount of resources used.

3. As can be seen in Figure 4, its central part, CUDA SRK kernel is invoked within the main loop, running in CPU, and mapped on threads configuration. This main loop is governed by the number of integration steps. It means, 10^8 times or more. We were able to achieve a significant improvement in overall system performance. The timed computation is represented in the central part of the figure, including both CPU main loop and memory transfer overhead between memory RAM and GPU device. Chain oscillator interactions as we see in Figure 2 represent a problem to solve. So, we used shared memory into kernel code when accessing data structures to improve performance. The maximum speedup achieved is 15X for 8192 oscillators, and it was measured as the ratio between serial and parallel time. Serial run was done on a CPU Intel 8 Core i7-2600 - 3.4 GHz. Kernels execution time remain constant between 16 to 256 oscillators, then between 1024 and 8192 oscillators.The warps threads don't diverge and they keep maximum GPUs

occupancy. In Figure 6, we show this results when the kernel is launched with 256 threads per block. As can be seen, we increase the load, but time does not increase, and we can process 8192 oscillators at the same time as 2048!, meanwhile CPU execution time grows exponentially with load. For this reason, GPU approach is a very good choice.

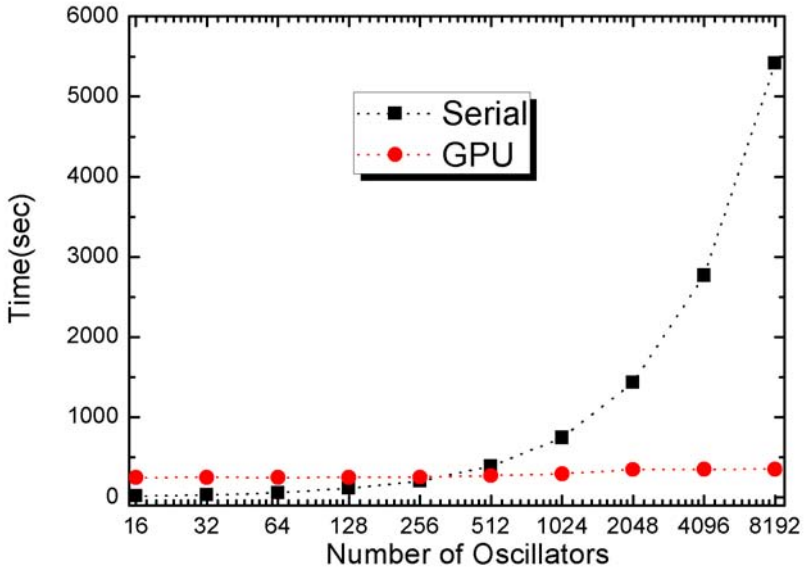


Fig. 6. GPU and Serial runtimes. Results represent execution time, using Time Steps=8000000 and Threads per Block=256.

5 Conclusions and Future Work

This work has applicability in the modelling of many low-dimensional physical systems and technological applications in different scales. In particular our proposal is the study of the heat transport phenomenon along one-dimensional nanodevices. Despite the model studied here is rather simple, it is usually a good first approach to achieve a qualitative and quantitative physical insight of the phenomenon of energy transfer. On the other hand, the computational model of this problem consumes a high runtime, prompting the use of parallel computing resources.

We have shown the suitability of commodity GPUs and a parallel CUDA-based algorithm for solving this problem, in particular when long chains are considered. The MPI implementation is inefficient in this case. We are working to further improve these results taking advantage of the availability of hibrid computers for high performance computing.

6 Acknowledgment

The authors thank to CECAR - FCEN - UBA and III-LIDI- Facultad de Informtica - UNLP for the opportunity to have access to a 32-Core IntelXeon and a 8-Core Xeon in CECAR and a BLADE multicore Xeon(TM) 5030 and a GPU Geforce GTX 560TI in III-LIDI.

References

1. Beraha N., Barreto R., Soba A., Carusela M.F. in preparation.
2. Dongarra J., Sterling T., Simon H., Strohmaier E., High-performance computing: clusters, constellations, MPPs, and future directions. Journal of Computing in Science & Engineering, **7** (2005) 51–59
3. Honeycutt, Rebecca L., Physical Review A (Atomic, Molecular, and Optical Physics), Volume 45, Issue 2, January 15, (1992), pp.600-603
4. Januyszewski M., Kostur M., Accelerating numerical solution of stochastic differential equations with CUDA. Computer Physics Communictions. Elsevier. **181** (2010) 183–188
5. Jin H., Jespersen D., Mehrotra P., Biswas R., Chapman B. High performance computing using MPI and OpenMP on multi-core parallel systems. Parallel Computing. Elsevier. **37** (2011) 562-575
6. Lobachev O., Guthe M., Loogen R. Estimating parallel performance. Journal of Parallel and Distributed Computing. Elsevier. **73** (2013) 876-887
7. Nickolls J., Dally W.: The GPU Computing Era. IEEE Micro. (2010) 56–69
8. Muresano R., Rexachs D., Luque E. How SPMD applications could be efficiently executed on multicore environments?. IEEE International Conference on Cluster Computing and Workshops. (2009)
9. Schenk O., Christen M., Burkhart H.: Algorithmic performance studies on graphics processing units. Parallel Distributed Computing. Elsevier. **68** (2008) 1360–1369

Desarrollo de aplicaciones paralelas en Erlang/OTP utilizando múltiples *planificadores*

Juan Ernesto Pisani[1], Pablo Cristian Tissera[1], A. Marcela Printista[1,2]

[1] Departamento de Informática, Laboratorio de Investigación y Desarrollo en Inteligencia Computacional. UNSL.
[2] Conicet - CCT San Luis.
{juanpisani, ptissera, marprinti}@gmail.com

Resumen.
Este trabajo se enfoca en el desarrollo de aplicaciones distribuidas y paralelas mediante el uso de un lenguaje orientado a la concurrencia, denominado Erlang. El objetivo es explorar las capacidades y limitaciones de este lenguaje de programación cuando se desea implementar aplicaciones sobre arquitecturas multicore. Como caso de uso, se describe una secuencia de implementaciones de un modelo de Autómata Celular, el cual presenta ciertas características específicas que lo hacen un modelo atractivo para aplicación de técnicas de paralelización.

Palabras claves: leguaje Erlang, máquina virtual, soporte multicore, planificadores.

1 Introducción

La primera implementación del lenguaje Erlang [1,7], en la década del 70, consistió de un intérprete implementado en Prolog [3,4] para el álgebra desarrollada por la Empresa de Telefonía Ericsson. Dos décadas mas tarde, con la incorporación del entorno de ejecución denominado, *BEAM Virtual Machine,* logra su aceptación en la comunidad de sistemas distribuidos.

Erlang es considerado un **lenguaje de programación orientado a la concurrencia**. Además, Erlang se ajusta en muchos aspectos al **paradigma de lenguaje funcional**. Al igual que los demás lenguajes funcionales enfatiza la aplicación de funciones en contraste con los cambios de estados típicos del paradigma imperativo. Erlang resalta las principales ventajas de todo lenguaje funcional como por ejemplo la ausencia de efectos colaterales, la mayor facilidad para la depuración de programas y la mayor facilidad para la ejecución concurrente, entre otras.

Sumado a estas características, el lenguaje también presenta propiedades de un paradigma de orientación a objetos, como es el caso del polimorfismo el cual admite sobrecargar una función determinada en base a la cantidad de parámetros que la misma reciba en tiempo de ejecución. Actualmente existe una discusión abierta concerniente a todas las características que el lenguaje presenta y que lo apartan de la

definición pura de lenguaje funcional. Por ello a menudo se lo cataloga como lenguaje *híbrido*, dado que un subconjunto del mismo es estrictamente funcional pero incluye construcciones del tipo imperativo e incluso rasgos de la orientación a objetos, aunque no respeta completamente ninguno de estos paradigmas.

Es relevante mencionar que como lenguaje orientado a la concurrencia Erlang proporciona facilidades para la programación distribuida y paralela. El lenguaje integra un conjunto sencillo pero completo de operaciones que permiten la creación y manejo de procesos concurrentes dentro de su máquina virtual. Erlang utiliza procesos ligeros cuyos requisitos de memoria pueden variar de forma dinámica. Los procesos no tienen memoria compartida y se comunican por paso de mensajes asíncronos. En un entorno distribuido, la *Máquina Virtual Erlang* recibe el nombre de Nodo Erlang, pudiendo llegar a constituirse una red de nodos Erlang, cuyos procesos se comunican mediante el paso de mensajes. La administración del hardware subyacente se mantiene en forma transparente a la resolución del problema, a diferencia de lenguajes de programación como C que utilizan librerías auxiliares como PVM o MPICH para lograr el mismo propósito.

En 2010 Erlang agregó soporte SMP (acrónimo del inglés *Symmetric Multi-Processing*) y más recientemente soporte para plataformas multicore, características que serán ampliadas en la sección 2 del presente trabajo.

Este trabajo tiene como objetivo presentar el desarrollo de una aplicación distribuida utilizando el lenguaje Erlang y analizar la performance de la misma sobre computadoras que disponen de múltiples cores para su ejecución. En la sección 3 se define nuestro caso de estudio que consiste de la Simulación de un autómata celular, la sección 4 desarrolla las implementaciones realizadas y la sección 5 realiza un análisis comparativo de las mismas. Finalmente la sección 6 presenta las conclusiones de esta experiencia y el trabajo de investigación futuro.

2 Modelo de Programación Concurrente

En Erlang toda actividad concurrente es encapsulada dentro de una entidad computacional denominada proceso. Los procesos en Erlang son entidades independientes, no cuentan con memoria compartida ya que cada proceso tiene su propia memoria y los cambios realizados en la misma son de carácter local. Los distintos procesos interactúan entre ellos vía pasaje de mensajes asincrónicos.

Cabe aclarar que en este contexto haremos uso del término *proceso* para referirnos a entidades independientes dentro del entorno de ejecución de Erlang. Dichos procesos son administrados directamente por el entorno de ejecución y no por el sistema operativo. Es importante también diferenciar el concepto de procesos Erlang, los cuales representan entidades independientes entre sí, con un thread de ejecución los cuales comparten estructuras de datos y recursos comunes. Generalmente los programas en Erlang se componen de decenas, miles o incluso cientos de miles de pequeños procesos, todos ellos funcionando de forma independiente e interactuando entre sí mediante el envío de mensajes para coordinar las distintas actividades a realizar y poder llegar a la solución de un problema determinado. Los programas organizados de esta manera escalan fácilmente ya que

ante una demanda excesiva de trabajo es posible crear más procesos de modo tal que la demanda sea satisfecha sin pérdida de rendimiento.

2.1 Programación Multicore

Como mencionamos anteriormente, en Erlang toda concurrencia es explicita y forma parte del lenguaje. El lenguaje brinda un conjunto reducido pero robusto de primitivas para poder crear procesos y facilitar las comunicaciones entre ellos. Este conjunto sólo incluye tres operaciones que permitirán en definitiva crear procesos, enviar mensajes a procesos previamente creados y recibir mensajes respectivamente.

Erlang provee soporte para una implementación Multicore. Esto se logra con una operación combinada entre la Máquina Virtual Erlang la cual proporciona el marco en el cual las aplicaciones serán ejecutadas y el soporte SMP provisto por el lenguaje. Las aplicaciones Erlang se compilan y luego son ejecutadas por la Máquina Virtual (es decir su entorno de ejecución) independientemente del hardware subyacente que se utilice.

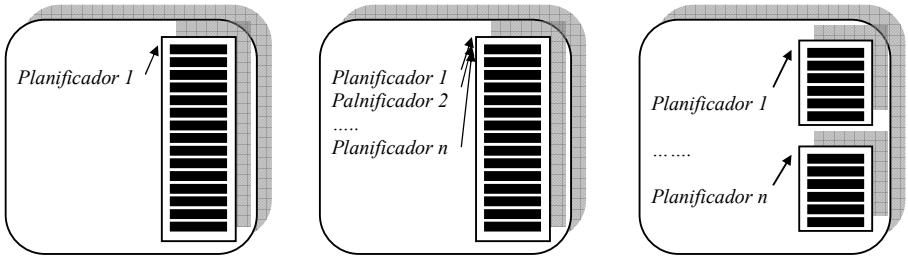


Fig. 1. Máquina Virtual Erlang: (izq) sin Soporte SMP (centro) varios panificadores para coordinar el trabajo desde una única cola y (der) varios planificadores organizan el trabajo desde n colas de trabajos.

Inicialmente, el soporte SMP se encuentra deshabilitado, por lo que el entorno de ejecución de Erlang inicia solo un planificador el cual se ejecuta sobre el thread principal de ejecución de la Máquina Virtual. La función de este planificador es coordinar directamente los distintos procesos y trabajos de entrada/salida, a medida que los mismos se encuentren disponibles, desde una única cola de ejecución como puede verse en la Fig. 1 (izq).

En el caso de hacer uso del soporte SMP, se podrán iniciar varios planificadores, los cuales se ejecutarán sobre threads a nivel de aplicación totalmente independientes, situación que se puede observar en la Fig. 1 (centro). Se debe tener en cuenta que los distintos planificadores, en este caso, estarán bajo una condición de competencia permanente requiriéndose mecanismos de sincronización para el acceso a la cola de ejecución. Esta implementación en particular fue utilizada para las primeras versiones de la plataforma que brindaron soporte SMP para arquitecturas de hardware multicore.

Para solucionar el problema de sincronización se introdujeron múltiples colas de ejecución, característica útil en caso de disponer de varios recursos de computación (cores). Como se puede observar en la Fig. 1 (der) los procesos son asignados a distintos planificadores y cada uno de ellos podrá planificar su ejecución de forma independiente.

Finalmente se introdujo lógica de migración de procesos entre planificadores a fin de evitar que la proliferación de procesos sobrecargue alguna cola de ejecución en tiempo de ejecución. La lógica de migración existente debe ser eficiente y a la vez razonablemente justa para garantizar un balance de carga de los distintos procesos entre los planificadores instanciados. Esta configuración se muestra en la Fig. 2.

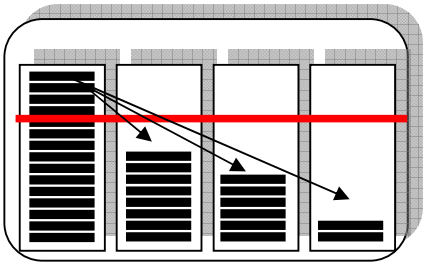


Fig. 2. Máquina Virtual Erlang con soporte SMP, varios planificadores, varias colas de ejecución y con capacidades de migración.

La lógica de migración implementada resulta transparente para el programador.

3 Caso de Estudio: Simulación del Juego de la Vida

El autómata celular (AC) [4] que resuelve *El Juego de la Vida* fue diseñado por el matemático británico John Horton Conway en 1970. Su nombre se debe a la estrecha analogía que tiene el juego con el auge, la caída y alternancias de una sociedad de organismos vivos. Desde su publicación [5], ha atraído mucho interés debido a la gran variabilidad de la evolución de los patrones.

Para la simulación del Juego de la Vida, se utiliza un modelo de autómata celular, el cual consiste de una cuadrícula dividida en celdas contiguas que representará el tablero de juego. Para la evolución de la simulación, cada celda chequea el valor de su celda y el de sus ocho celdas vecinas (vecindad de Von Neuman) y entonces decide acerca de la vida o muerte de un organismo en la misma. Cada celda del tablero contiene a lo sumo un organismo y los distintos organismos residentes en las distintas celdas son los participantes del juego. Su evolución está determinada por el estado inicial del juego y no requiere de ninguna entrada de datos posterior.

La configuración inicial del juego consiste en distribuir una determinada cantidad de organismos sobre las distintas celdas disponible y a partir de allí observar como las "leyes genéticas" (definidas por Conway) se aplican para los nacimientos,

muertes y supervivencias de los distintos organismos o población inicial que participa en el juego:

1. Supervivencia: cada organismo rodeado por dos o tres organismos en la presente generación, sobrevive para la próxima generación.
2. Nacimiento: toda celda sin un organismo y rodeada por tres organismos en la presente generación, incluirá un organismo en la próxima generación.
3. Extinción (Disgregación): cada organismo rodeado por sólo un organismo o ninguno en la presente generación, morirá en la próxima generación.
4. Extinción (Hacinamiento): cada organismo rodeado por cuatro o más organismos en la presente generación, morirá para la próxima generación.

En un AC, las reglas de evolución se aplican simultáneamente a todas las celdas del autómata, lo que constituye una generación. En este caso de estudio, la aplicación de las "leyes genéticas" a todas las celda del tablero, constituye un "movimiento" en la "historia de vida" del juego. Por ende la primera generación se obtiene luego de aplicar las reglas de evolución a la configuración inicial del juego. A partir de allí podremos repetir el procedimiento para obtener sucesivas generaciones. Durante el transcurso del juego se puede observar cómo la población inicial evoluciona constantemente debido a cambios que resultan inesperados, aunque esto puede no ocurrir hasta después de un gran número de generaciones. La Fig. 3. muestra ejemplos de patrones particulares en los cuales se pueden observar distintos comportamientos generalizados para el juego de la vida.

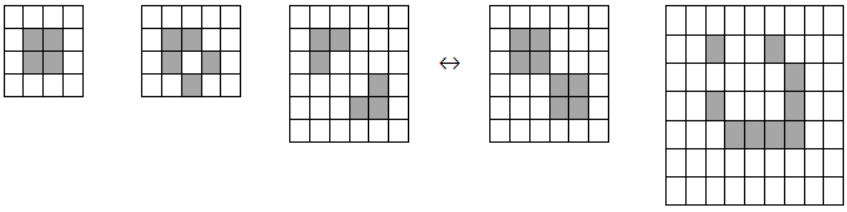


Fig. 3. Patrones del Juego de la Vida: estáticos (1ero. y 2do.), recurrente (3ro. y 4to.) y un patrón infinito (5to.)

4 Implementaciones

Para el desarrollo de la aplicación concurrente que describiremos en este trabajo, se utilizó el framework llamado OTP (acrónimo del inglés *Open Telecom Platform*) [6,8]. Esta herramienta de trabajo organiza los programas en una jerarquía de procesos conocida como *árbol de supervisión*. Dicha jerarquía estructura los distintos procesos que conforman un programa en procesos **Subordinados o Workers,** aquellos procesos que se encargan de procesar la lógica interna de un programa determinado y los **Supervisores** que controlan el comportamiento de los workers, los inician o finalizan. Los procesos supervisores y workers cuentan con una estructura

similar y la única diferencia existente entre ellos es la posibilidad, por parte de los primeros, de poder reiniciar a aquellos procesos que supervisan.

Para la experimentación se consideró un autómata celular inicializado aleatoriamente y un periodo de evolución igual a 100 generaciones. El espacio celular del autómata está representado por una matriz bidimensional NxN donde N=256. Los test correspondientes fueron ejecutados sobre un multiprocesador de 32 cores, con memoria de 64GB (16x4GB), 1333MHz y un disco de 500GB. 4x AMD Opteron 6128, 2.0GHz, 8C, 4M L2/12M L3, 1333 Mhz. El sistema operativo utilizado es Linux Centos 6 con soporte de 64 bits. Los resultados se obtuvieron variando el número de planificadores o threads de la Máquina Virtual Erlang y asignando cada uno de ellos a un core distinto del servidor multicore disponible.

Primer Implementación (v.1)

En esta implementación *cada celda fue representada por un proceso independiente*, de modo tal que el conjunto de procesos representa el espacio celular del autómata. El tablero de juego puede visualizarse como un conjunto de procesos interconectados entre sí dentro del entorno de ejecución de Erlang. Esta implementación requirió de un proceso adicional para sincronizar la evolución general del autómata, quién también hizo las veces de un servidor centralizado de búsqueda (CSS del acrónimo *Central Search Server*) encargado de identificar a la vecindad de un proceso.

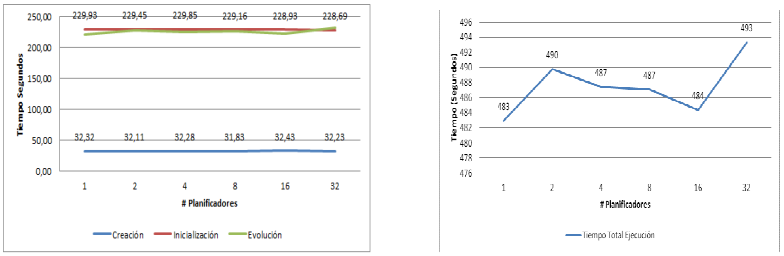


Fig. 4. v.1: (izq) tiempo total de creación de procesos, de inicialización de procesos y total de evolución; (der) tiempo total de ejecución de la simulación.

Como puede observarse en la Fig. 4, los resultados de la v.1 distan de ser alentadores en el contexto de computación distribuida o paralela, dado que los tiempos de ejecución de la aplicación no varían al aumentar el número de planificadores que utilice el entorno de ejecución de Erlang.

Segunda Implementación (v.2)

El componente CSS utilizado en la v.1 utiliza como principal estructura de datos para almacenar toda la información relacionada a los procesos, una lista secuencial ordenada. La alta demanda de trabajo requerido por todos los procesos creados hace que esta lista secuencial ordenada sea una mala opción como estructura de datos inicial dado que limita severamente el rendimiento del sistema completo.

Con el objetivo de acelerar el proceso de inicialización se reemplazó la lista secuencial por una estructura denominada *ETS* (acrónimo del inglés, *Erlang Term Storage*) la cual forma parte del lenguaje y cuenta con soporte para operaciones de lectura/escritura concurrente. Conceptualmente una tabla ETS es una estructura de datos capaz de almacenar cualquier término *clave – valor* en Erlang, en una analogía directa con arreglos asociativos encontrados en muchos otros lenguajes de programación. Erlang provee un conjunto amplio de operaciones que pueden realizarse con tablas ETS, como por ejemplo consultas que involucren *pattern-matching*. El tipo determinado de una tabla ETS queda establecido al momento de su creación [1].

Para almacenar la información relacionada a los procesos del autómata, la v.2 utiliza una tabla ETS del tipo *ordered_set*. Entre la información que se almacena en esta estructura se encuentra la posición determinada (X, Y) de cada proceso, la cual se utiliza como clave principal en la tabla ETS, y el Identificador de Proceso o PId el cual será compartido por los distintos procesos que lo requieran.

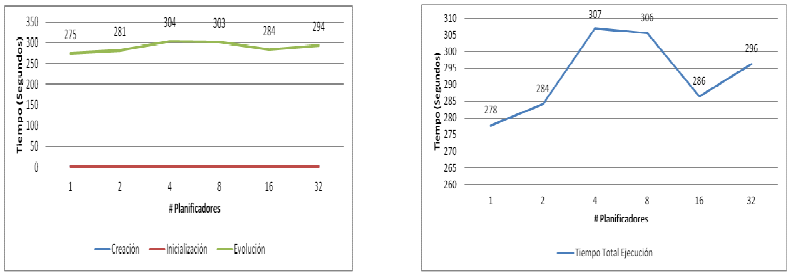


Fig. 5. v.2: (izq.) tiempo total de creación de procesos, de inicialización de procesos y de evolución; (der.) tiempo total de ejecución de la simulación.

La Fig. 5 (izq.) refleja la reducción en el tiempo total requerido para completar la fase de inicialización producto del uso de una estructura más eficiente. Sin embargo, se puede observar que el tiempo de simulación menor es logrado con un único planificador. Este pobre desempeño, se debe a probables cuellos de botella que permanecen y que serán abordados en próximas versiones.

Tercera Implementación (v.3)

La coordinación existente entre el proceso administrador y los demás procesos workers produce una merma en el rendimiento de la aplicación ya que todo el procesamiento se concentra sobre un único punto de control. Durante el tiempo que le insume al proceso administrador distribuir todo el trabajo entre los procesos existentes, el resto de los planificadores permanecen ociosos. En este escenario, solo un único planificador dispone de exclusividad para ejecutar su código fuente mientras los demás planificadores permanecen ociosos hasta que la ejecución del proceso administrador termina.

La v.3 replantea la estructura jerárquica utilizada y propone utilizar tantos procesos administradores como planificadores se estén utilizando en el entorno de ejecución de Erlang. De este modo se espera adaptar la ejecución del programa de acuerdo a la disposición de hardware con la que se cuenta.

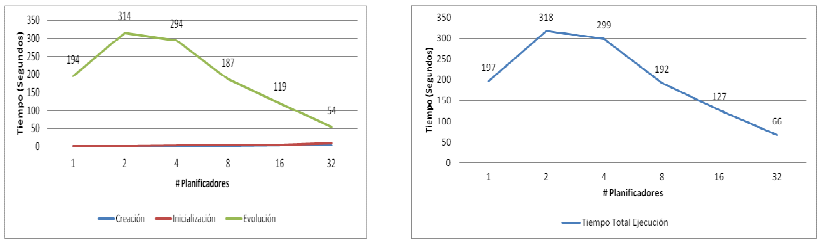


Fig. 6. v.3: (izq.) tiempo total de creación de procesos, de inicialización de procesos y total de evolución; (der.) tiempo total de ejecución de la simulación.

Como se muestra en la Fig. 6, la nueva estructura de la aplicación con múltiples procesos administradores hace que el rendimiento de la aplicación mejore a medida que se disponen de más planificadores durante el tiempo de ejecución.

Cuarta Implementación (v.4)

Esta implementación tiene el objetivo de maximizar la performance de la aplicación evitando penalidades relacionadas a una excesiva administración de procesos. La nueva estructura de programa consta de procesos workers capaces de contener más de una celda del autómata celular en su estado interno, quedando el espacio celular distribuido equitativamente entre los procesos creados inicialmente. Toda esta información será almacenada en una tabla ETS a fin de que las distintas operaciones que deban realizarse sobre ella sean eficientes.

El código fuente del programa se simplificó dado que en esta implementación no se utiliza un proceso como CSS. Tampoco es requerida una etapa de inicialización de procesos ya que las distintas tablas ETS creadas utilizan como clave de búsqueda la posición (X, Y) de una celda dada del autómata.

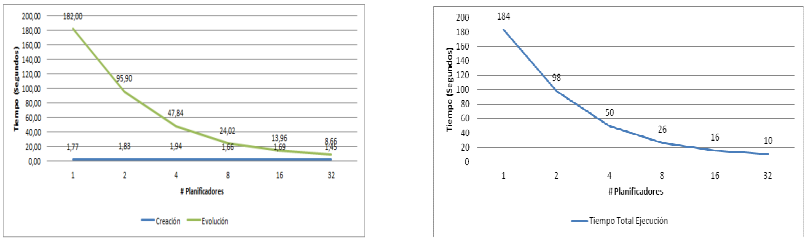


Fig. 7. v.4: (izq.) tiempo total de creación de procesos, de inicialización de procesos y total de evolución; (der.) tiempo total de ejecución de la simulación.

Los resultados mostrados en la Fig. 7. reflejan que una reducción en el número de procesos creados inicialmente para resolver la aplicación conlleva una menor demanda de trabajo por parte del entorno de ejecución a la hora de administrar los distintos procesos creados durante la ejecución del programa, mejorando significativamente los tiempos totales de ejecución a medida que se agregaron planificadores.

5. Análisis Comparativo

Una observación inmediata surge al comparar las últimas dos versiones implementadas con Erlang/OTP, donde claramente el número de procesos creados en tiempo de ejecución representa un factor determinante a la hora de medir la performance de nuestras aplicaciones.

La v.3 utilizó tantos procesos como celdas existían en el espacio celular más un conjunto de procesos auxiliares. Este escenario demandó una mayor planificación de procesos para su ejecución por parte de los distintos planificadores instanciados. Una excesiva planificación implica cambios de contexto de procesos más frecuentes para permitir la ejecución de la mayor cantidad de procesos mientras se esté utilizando algún core o CPU físico.

Todas estas latencias involucradas durante la ejecución del programa resultaron en una performance pobre y sin posibilidad de escalar. La Tabla 1 compara el rendimiento de los tiempos de ejecución obtenidos utilizando un número creciente de planificadores. En el mejor de los casos, 32 planificadores, sólo se presentó un factor de mejora de 3 veces.

Tabla 1. Factor de rendimiento para la implementación v.3.

#Planificadores	Tiempo Total de Ejecución	Factor de Rendimiento
1	197 seg.	1X
2	**318 seg.**	**0,61X**
4	**299 seg.**	**0,65X**
8	192 seg.	1,02X
16	127 seg.	1,55X
32	66 seg.	2,98X

La v. 4, utilizó un enfoque de evolución distinto. La utilización de un número reducido de procesos condujo a un menor tiempo de latencia en planificación de procesos. Menor tiempo de latencia implica mayor tiempo para ejecución de procesos mientras se esté utilizando algún core específico lo cual impacta directamente sobre los distintos tiempos de ejecución del programa.

La Tabla 2 mide el rendimiento de la última implementación. Como se puede observar, disponer de mayor cantidad de planificadores conduce a una disminución en el tiempo total de ejecución, pero la eficiencia puede verse afectada considerablemente si la aplicación no realiza un balance entre número de planificadores y el número de cores disponibles.

Tabla 2. Factor de rendimiento para la implementación v.4

#Planificadores	Tiempo Total de Ejecución	Factor de Rendimiento	Rendimiento/ #Planificadores
1	184 seg.	1X	100%
2	98 seg.	1,88X	93,88%
4	50 seg.	3,68X	92,00%
8	26 seg.	7,08X	88,46%
16	16 seg.	11,50X	71,88%
32	10 seg.	18,40X	57,50%

6 Conclusiones

En los últimos años, paulatinamente han comenzado a recibir atención propuestas de lenguajes orientados explícitamente a la programación concurrente y paralela, tal es el caso de Erlang y su entorno de desarrollo OTP. En este trabajo se discutieron una serie de implementaciones distribuidas enfocándonos principalmente en el soporte SMP que el lenguaje provee para poder ejecutar en arquitecturas multicore.

La experimentación realizada en este trabajo, intentó mostrar como los distintos factores involucrados en un programa Erlang impactan en el rendimiento obtenido, donde evidentemente, el número de procesos o planificadores juega un rol fundamental.

El trabajo realizado en este paper, será transferido al desarrollo de aplicaciones distribuidas más complejas. El uso intensivo de procesos para mantener ocupados durante el mayor tiempo posible a los distintos cores de un procesador multicore, junto a la eliminación de cuellos de botellas secuenciales y efectos colaterales deberán ser tenidos en cuenta al inferir el comportamiento de estas aplicaciones.

Referencias

1. Armstrong, J.: *Programming Erlang: Software for a Concurrent World* Pragmatic Bookshelf (2007).
2. Kowalski, R.A.: *Logic for Problem Solving*, North Holland (1979)
3. Sterling, L., Shapiro E.: *The Art of Prolog*, MIT Press, (1986).
4. Wolfram, S.: *Cellular Automata and Complexity* (collected papers). Addison Wesley. (1994).
5. Gardner, M.: *Mathematical Games: The fantastic combinations of John Conway's new solitaire game "Life"*. Scientific American 223, 120–123 (1970).
6. Logan, M., Merritt E., Carlsson, R.: *Erlang and OTP in Action* – Manning Publications; Edición: Pap/Psc. (2010).
7. Erlang Programming Language: http://www.erlang.org
8. OTP Design Principles User's Guide: http://www.erlang.org/doc/design_principles/users_guide.html

Parameters Calibration for Parallel Differential Evolution based on Islands*

María Laura Tardivo[1,2,3], Paola Caymes-Scutari[2,3], Miguel Méndez-Garabetti[2,3] and Germán Bianchini[2]

[1] Departamento de Computación, Universidad Nacional de Río Cuarto. (X5804BYA) Río Cuarto, Córdoba, Argentina
lauratardivo@dc.exa.unrc.edu.ar

[2] Laboratorio de Investigación en Cómputo Paralelo/Distribuido (LICPaD) Departamento de Ingeniería en Sistemas de Información, Facultad Regional Mendoza Universidad Tecnológica Nacional. (M5502AJE) Mendoza, Argentina
{pcaymesscutari,gbianchini}@frm.utn.edu.ar
miguelmendezgarabetti@gmail.com

[3] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Abstract. We are studying different alternatives to obtain a version of the Differential Evolution (DE) algorithm that improves the solutions quality properties. One of the parallel alternatives, named Island Model, follows a Master/Worker scheme. With this model, multiple instances of DE are executed in parallel on various computing nodes or *islands*, each of them considering a different population of individuals. Each worker makes the search process, and communicates with each other to exchange information with certain frequency. This model significantly promote the exploration of a larger search space, which leads to good solutions quality. The aim of this paper is to analyse the behaviour of this model, when setting each island with different input parameters. We apply some input configuration tests for the islands, in order to analyse the impact in the solutions quality and the execution time, taking into account the crossover probability and mutation factor, and the crossing type. These parameters are crucial to guide the search towards certain areas of the search space.

1 Introduction

The interest in solving combinatorial optimization problems has gained popularity between the scientific and industrial community [15] [4]. Between the strategies developed to solve these problems we find specific heuristics and metaheuristics as popular techniques. The specific heuristics are problem dependent and are designed in a particular way to solve a given problem. Meanwhile, metaheuritics represent a more general set of solutions that can be applied to a large

* This work has been supported by UTN under projects PICT2010/12 and UTN1585, and by ANPCyT under project PRH PICT-2008-00242.

number of problems. Metaheuristics try to solve instances of the problem, exploring the wide space of solutions that those instances can admit [10]. Usually, these solutions are called optimum, in reference to the better or best values found for the optimization problem involved, leading to local (better solutions) or global (the best solution) optimum. Although obtaining the global optimum is desirable, sometimes the function to optimize is sufficiently complex to find the desired value within a reasonable time. For this reason, obtaining good quality "local optima" becomes a valid alternative.

The Differential Evolution (DE) algorithm is a population based metaheuristic, capable of working reliably in nonlinear and multimodal environments [9]. It starts to explore the search space by initializing multiple, randomly chosen initial points distributed in D-dimentional vectors that represent the individuals of the population. The algorithm presents two operators responsible for the creation of new solutions. First, the mutation operation creates a trial vector as a linear combination of some members of the population, then the crossover operation combines the parameter values of the trial vector with those of another member of the population, resulting in the target vector.

The classic version of DE follows a sequential processing scheme. However, the Differential Evolution algorithm (and in general metaheuristics) are naturally prone to parallelism, because most variation operations can be undertaken in parallel [1]. There are several studies that incorporate parallelism to DE, in order to improve the quality of the solutions obtained and/or diminish the execution time. In this work we follow the first objective. With the aim of improving the quality of solutions by exploring a larger sample domain, we focus on the *Island Model* [10]. Even though the execution time is similar to the sequential one, each island in the model is responsible for the evolution of the population that manages, and may use different parameter values and different strategies for any search component such as selection, replacement, variation operators (mutation, crossover), and encodings. An appropriate choice for its values may achieve quality and/or performance improvements.

In this work we present a study on the calibration for some parameters of the parallel *Island Model*, through the experimental study with various input configurations for each island of the model. The aim is to analyse the impact produced by these configurations, taking into account the quality of the solutions and the execution time of the parallel algorithm. Specifically, we focus on three of the most important input parameters of DE. They are the crossover probability and mutation factor, and the crossover type. It is known that the choice of their optimal values is an application dependent task. For two optimization benchmark problems under study, we want to get an overview about the behaviour of this model applied to solve them, considering different scenarios.

The paper is organized as follows: Section 2 describes the main characteristics of DE. Section 3 present a complete description of the *Island Model* used in this work, including its main features and its processing scheme. Section 4 shows the experiments carried out and the analysis of results. Finally, we present the conclusions and future work.

2 Classical Differential Evolution

The Differential Evolution algorithm has emerged as a popular choice for solving global optimization problems . Using a few parameters, it exhibits an overall excellent performance for a wide range of benchmark as well as real-world application problems [2]. Each individual belongs to a generation g, i.e., let $X_{i,g} = (x^1_{i,g}, ..., x^D_{i,g})$ an individual of the population, with $i = 1, ...,$N where the index i denotes *i-th* population individual and N is the total number of individuals in the population. Following, we explain the three classic main operators of DE. In section 3 we also introduce the *migration* operator, which is typically used in parallel versions of metaheuristics.

Mutation: After initialization, DE mutates and recombines the current population to produce another one constituted by N individuals. The mutation process begins in each generation selecting random individuals $X_{r_1,g}, X_{r_2,g}$. The *i-th* individual is perturbed using the strategy of the formula (1), where the indexes i, r_1 and r_2 are integers numbers different from each other, randomly generated in the range [1, N].

$$\text{“}DE/best/1\text{”} : V_{i,g+1} = X_{best,g} + (X_{r_1,g} - X_{r_2,g})F \tag{1}$$

The constant F represents a scaling factor and controls the difference amplification between individuals r_1 and r_2. It is used to avoid stagnation in the search process. $X_{best,g}$ is the best individual, i.e., it has the best value of the objective function evaluation among all individuals of current generation g. The notation “$DE/best/1$” represents that the base vector chosen is the best individual, and “1” vector difference is added to it.

***Crossover*:** After the mutation phase, each perturbed individual $V_{i,g+1} = (v^1_{i,g+1}, ..., v^D_{i,g+1})$ and the individual $X_{i,g} = (x^1_{i,g}, ..., x^D_{i,g})$ are involved in the crossover operation, generating a new vector $U_{i,g+1} = (u^1_{i,g+1}, ..., u^D_{i,g+1})$, denominated “trial vector”, and obtained using the expression (2).

$$U^j_{i,g+1} = \begin{cases} v^j_{i,g+1} & \text{if } rand_j \leq Cr \text{ or } j = k \\ x^j_{i,g} & \text{in other case} \end{cases} \tag{2}$$

where $j = 1, ..., D$, and $k \in \{1, ..., D\}$. The latter is a randomly generated index chosen for each individual. This index is used to ensure that the trial vector is not exactly equal to its source vector $X_{i,g}$, then a vector component at position k is taken from the mutated vector. The constant Cr, denominated *crossover factor*, is a parameter of the algorithm defined by the user. Cr belongs to the range $[0, 1]$ and is used to control the values fraction that are copied from the mutant vector V. $rand_j$ is the output of an uniformly distributed random number generator, and is generated for each component $U^j_{i,g+1}$ of the trial vector.

There are two crossing operators that can be applied: binomial or exponential. Both types use the expression (2), but differ in the way it is applied. The binomial crossover operator iterates over all the components of the individual, copying the jth parameter value from the mutant vector $V_{i,g+1}$ to the corresponding element in the trial vector $U_{i,g+1}$ if $rand_j \leq Cr$ or $j = k$. Otherwise,

it is copied from the corresponding target (or parent) vector $X_{i,g}$. Instead, the exponential crossover operator inherits the parameters of trial vector $U_{i,g+1}$ from the corresponding mutant vector $V_{i,g}$ starting from a randomly chosen parameter index, until the jth parameter value satisfying $rand_j > Cr$. The remaining parameters of the trial vector $U_{i,g+1}$ are copied from the corresponding target vector $X_{i,g}$.

***Selection*:** This phase determines which element will be part of the next generation. The objective function of each trial vector $U_{i,g+1}$ is evaluated and compared with the objective function value for its counterpart $X_{i,g}$ in the current population. If the trial vector has less or equal objective function target value (for minimization problems) it will replace the vector $X_{i,g}$ in the next generation. The scheme followed is presented in the expression (3).

$$X_{i,g+1} = \begin{cases} U_{i,g+1} & \text{if } f(U_{i,g+1}) \leq f(X_{i,g}) \\ X_{i,g} & \text{in other case} \end{cases} \tag{3}$$

The three stages mentioned above are repeated from generation to generation until the specified termination criterion is satisfied. This criterion could be finding a predefined minimal error or reaching a certain number of iterations.

Due to the potentialities provided by DE numerous variations and methods have been proposed with the aim of improving the performance of the classic technique. Among them are those trying to adapt the parameters of the algorithm, such as self-adjusting [14], [12], [3]; others using different mechanisms to optimize the individuals selection for the mutation and crossover phases [4], and some combining both methods [15]. In the next section we briefly mention some related works on parallel DE, and we present the details of our parallel proposal.

3 Island Parallel Model for Differential Evolution

Researchers have proposed different approaches to parallelize population-based metaheuristics, depending on the purpose to be achieved. In [16] is presented a proposal for solving the Pareto front problem. An individual in the population can be migrated with a certain probability to a random position in a random subpopulation. In [11], the model uses a ring interconnection topology and random migration rate controlled by a parameter of the algorithm. The aim of that work is to study the implications of a controlled migration constant. In [5], a parallel DE version is proposed and applied to solve biological systems. It also follows a ring interconnection topology. The analysis was done with different migration rates and they conclude by identifying the best of them. A critical issue of the last three approaches, is the consideration of the population size. The ability of the algorithm to find a solution depends on the tasks size and is related to the amount of individuals per node, making significant local and global evolution. Then, it is relevant to make experiments that encompass these scenarios. For all these reasons arises the need to perform a comparative study to test with large enough cases.

Following, we will describe the ***Island Parallel Model***. It follows a *Master-Worker* [6] scheme. Multiple instances of DE are executed in parallel on different computing nodes, each one considering a different population of individuals (Pop. 0, ..., Pop. n). We call each computing node "an island". A master process is in charge of monitoring the system as a whole, and each worker process is dedicated to compute all the generations in that island. Figure 1 represents this model. As can be seen, the master process is located in an exclusive computing node, so as to coordinate the system and to avoid delaying the response to the workers.

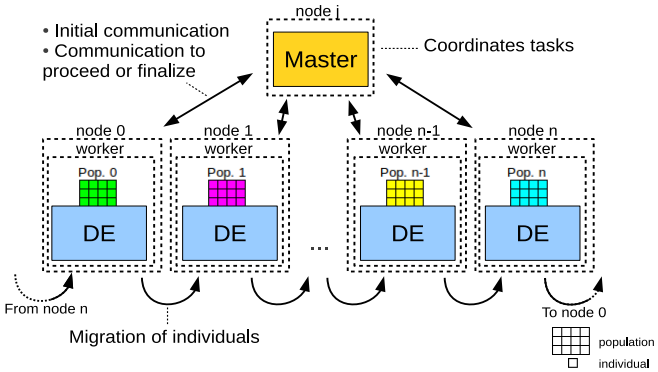


Fig. 1: **Island Model:** independent or cooperating self-contained metaheuristics.

Every certain number of generations, and considering a certain topology, begins a ***migration phase***. The amount of individuals that migrate is a certain percentage of the whole population, calculated from the number of individuals in the island. This *migration phase* represents another operator for parallel DE, and its importance lies in the need to exchange information between the islands to keep global information.

After a migration phase and replacement process, the workers inform to the master which is the best individual found. The master receives this information and temporarily stores the best individual of all those who have been sent by the workers. Then, if the termination condition is met, the master sends a message to workers indicating the end of the process. Otherwise, the master informs to continue with their evolutionary process. In our proposal, the finalization condition was defined as reaching a certain number of generations.

We recall that the Island Model significantly promote the exploration of a larger search space, because the workers explore different search spaces, since each population is initialized with a different seed. This leads to better solutions quality, although the execution time is generally higher than that of the sequential version. This parallelism technique, where multiple instances of an algorithm are launched in parallel and interrelated is useful when the aim is to deepen the search, with no particular requirements for reducing the execution time.

The following section will describe some experiments made with the aim of analysing the solutions quality and the execution time when introducing certain configurations for each island. The goal of this calibration is to adjust the effec-

tiveness of the model, considering each population with different configurations for the mutation factor and crossover probability, and the crossing type. These parameters are crucial to guide the search towards certain areas of whole search space. If these parameters are set in an inappropriate manner, it may happen that the algorithm get stagnated in a local optimum, or the solutions quality obtained may be non optimal.

4 Test cases and analysis of results

In the following, we describe the experiments carried out in order to test the Island Model with different configurations. In the experiments, the performance of the algorithm was tested with a set of scalable functions, obtained from [13]. For each of them, 30 executions were carried out with different seeds. The sizes of the problems considered have dimensions 100, 500 and 1000. The population was made up with 100 and 400 individuals. The function used for the test where Shifted Sphere (unimodal, search range in [-100,100], bias value of -450), and Shifted Rosenbrock (multimodal, search range in [-100,100], bias value of 390).

The average error is defined as the difference between the current value of the global optimum and the value obtained by the algorithm. If the error is zero indicates that it has been found the global optimum. For the problems considered, the best results are those that are closer to zero error.

Preliminary experiments carried out on the model conduced to a definition for the exchange rate value. In all the tests the individuals were exchanged among the islands at a migration rate of 15% every 500 iterations. It is known from literature [4], [15], [14] that the values F=0.5 and Cr=0.3 may guide the search towards good solutions. We validated and established those values in the tests.

Four experiments were carried out; some of them are associated to the crossover probability and mutation factor, and others are related to the crossover type. Although there exists a wide range of combinations and possibilities of variation on these parameters, carrying out the test and processing the results are time-consuming actions. Our test are performed with large enough dimensions, to contemplate complex optimization problems. This is an important feature, that differentiates our case of analysis regard to those cases treated in other similar studies (such as those referenced before).

Following, we provide a brief description of the test cases:

- **Case 1:** This experiment consisted in the configuration of all the islands with the same input parameters (i.e. just varying the initial seed for each island). The goal is to explore the space more thoroughly. All islands try to solve the whole optimization problem searching in a different area of the search space, having the same configuration for the rest of the parameters.

- **Case 2:** This experiment consisted in setting the half of the islands in the model with random values for the mutation and crossover probabilities, and the other half of the islands used the fixed constant values for F and Cr.

- **Case 3:** The third experiment was performed with the aim of represent an independent behaviour of the islands, setting the input probabilities with

random values for each island in the model. This randomized configuration may reproduce a realistic scenario when the model observed is similar to a concrete natural system, where each population have its own working method. In this sense, the complete problem is solved by different entities, having their own search space and a unique search configuration.

- **Case 4:** The last experiment involved the crossover type. All the islands were setted with the constant values for F and Cr. In this experiment we changed the crossover type to exponential crossover. With this case, we test the behaviour of the islands when the crossover type is distinct from the classic binomial one, verifying if it may conduct the search process towards other areas. This experiment can be contrasted against the Case 1.

The cases 1, 2 and 3, were performed using a binomial crossover type. Also, we can notice that the case 2 is middle point test between case 1 and case 3, trying to produce an hybrid scenario.

The islands follows a ring intercommunication topology, so that each island receives individuals from its predecessor in the topological order, and sends their own individuals to its successor in that order. The individuals to be migrated are the best member of the island plus other individuals randomly selected, and the received individuals will replace the worst members of the target population.

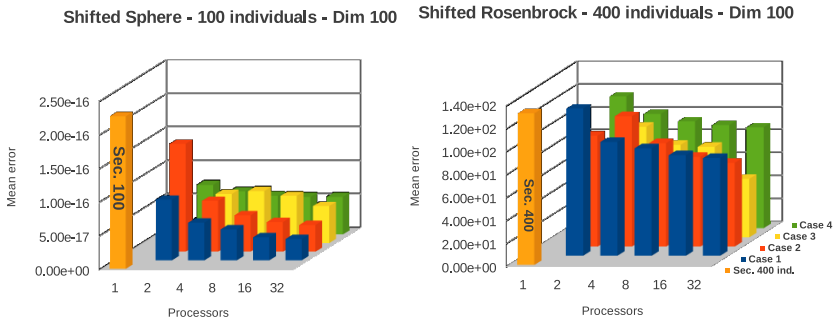


Fig. 2: Mean error of Shifted Sphere and Shifted Rosenbrock functions. Dimension 100.

In order to test scalability, all experiments included 2, 4, 8, 16 and 32 processors dedicated to the worker processes, and a separate processor for the master process. All tests were made on a cluster with 36 CPUs distributed between 9 nodes. They have 64 bits with Intel Q9550 Quad Core 2.83GHz processors and RAM memory of 4GB DDR3 1333Mz. All the nodes are connected together by Ethernet segments and switch Linksys SLM2048 of 1Gb. Base software on the cluster includes a 64 bits Debian 5 Lenny Operating System. In the codification we use the MPICH library [7] for message passing communication between participating nodes. Our algorithmic version of the Island Model is based on the sequential version of DE, obtained from [9].

Table 1 shows the average computing time, discriminating the tests according to the dimension and case analysed. The graphs of the figures 2, 3 and 4 show the mean errors obtained in the different experiments performed. In the graphs, each color represents one of the experiments mentioned above. In order

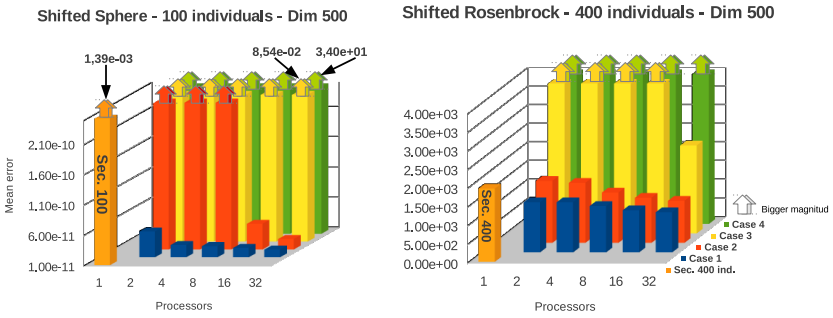


Fig. 3: Mean error of Shifted Sphere and Shifted Rosenbrock functions. Dimension 500.

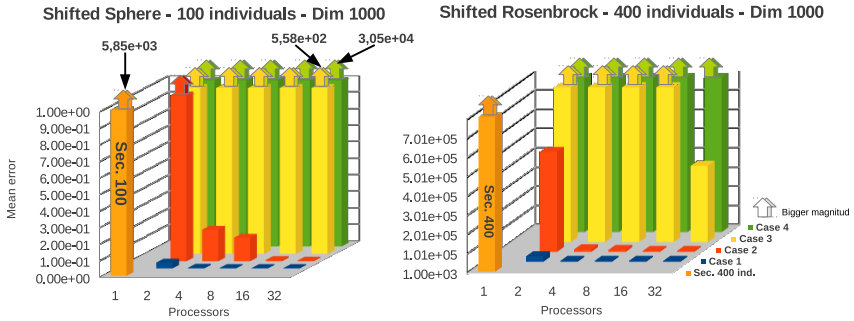


Fig. 4: Mean error of Shifted Sphere and Shifted Rosenbrock functions. Dimension 1000.

Table 1: Shifted Shpere and Shifted Rosenbrock average computing time (in seconds), obtained with the algorithm with each version.

Shifted Sphere 100 ind.						Shifted Rosenbrock 400 ind.					
Test case	2	4	8	16	32	Test case	2	4	8	16	32
Dim 100 Sequential time:=2.08						**Dim 100 Sequential time:**=12.82					
Case 1	2,68	2,71	2,83	2,91	4,27	Case 1	12,71	13,00	13,35	13,77	19,02
Case 2	2,78	2,73	3,00	3,05	3,09	Case 2	12,81	13,34	14,11	14,02	14,20
Case 3	2,81	3,01	3,04	3,04	3,11	Case 3	13,35	14,33	14,29	14,45	14,56
Case 4	0,41	0,42	0,42	0,48	0,57	Case 4	3,64	3,66	5,43	7,83	10,00
Dim 500 Sequential time:=11.06						**Dim 500 Sequential time:**=63.24					
Case 1	12,98	13,23	13,60	14,66	19,30	Case 1	64,39	82,11	84,42	91,83	109,11
Case 2	13,50	13,88	14,48	14,56	15,13	Case 2	69,44	84,61	92,98	95,07	99,47
Case 3	13,42	14,52	14,44	14,48	14,83	Case 3	74,75	87,86	84,23	98,77	97,46
Case 4	1,44	1,46	1,53	1,64	2,03	Case 4	18,84	23,17	22,80	26,40	32,78
Dim 1000 Sequential time:=22.57						**Dim 1000 Sequential time:**=128.64					
Case 1	26,03	26,85	28,18	29,37	40,10	Case 1	132,89	181,16	186,64	200,62	207,92
Case 2	27,08	28,47	29,11	30,08	39,48	Case 2	191,09	180,70	209,80	207,78	292,29
Case 3	28,61	30,60	30,67	34,01	32,77	Case 3	189,47	172,69	188,78	196,14	212,53
Case 4	3,12	3,25	3,53	5,17	7,69	Case 4	65,68	67,56	61,32	62,92	74,05

to contrast with the parallel experiments, the graphs also include two columns that represents the mean error for the sequential version. Some bar columns of the graphics have a colored arrow at top, representing that the column bar has a bigger magnitude that the maximum scale in the graphic. Moreover, we include some small labeled black arrows with the purpose of explicitly indicate the value of those big columns or to highlight some interesting value.

In first place we compare the results obtained for cases 1, 2 and 3. As can be seen, the execution time for them are similar. The test that obtains better quality results is the Case 1, i.e. the test in which all the islands are configured with the same values. When all the islands operate with the same diversification factors, the search is done in a better way. By contrast, when each island has a particular mutation and crossover probabilities, the results are not the best that can be achieved by the model. In second place, we compare cases 1 and 4. The case 4 obtains a significant reduction of the execution time. We recall that this case used the exponential crossover.

This type of crossover inherits the parameters of trial vector from the corresponding mutant vector, starting from a randomly chosen parameter index, until the jth parameter value satisfying $rand_j > Cr$. It is clear from this crossover type that when the condition $rand_j > Cr$ is met, the crossover iteration stops, so -in general- for each individual of the population, this action is less time consuming than the binomial crossover used in the rest of the experiments, in which all the vector components are involved. Frequently, this particularity leads to lower execution times in the overall process, but the quality of the solutions achieved is not the optimal. This can be one of the reasons because this crossover type is less used than the binomial one. But in some circumstances, it can be desirable to achieve less execution time relegating in some orders of magnitude the solutions quality. In such cases, the use of the exponential crossover can achieve that result. Then, in general terms, for these particular problems, setting both probabilities to constant values at model level and the crossover as the binomial one leads to better quality results.

5 Conclusions

In this paper we describe the Island Model used to obtain a parallel version for the Differential Evolution algorithm. Different experimental tests were carried out with the aim of analyse the behaviour of the model when each island is configured with different parameters. Our interest was on the crossover type and on the crossover and mutation probabilities, applied to solve the Shifted Sphere and Shifted Rosenbrock optimization problems. When using the exponential crossover type, the quality of the obtained solutions was not optimal. However this experiment achieved a significant reduction in execution time, because the crossover type characteristics. For this reason, the use of the exponential crossover may be useful when what is desired is a reduction in the execution time, relegating in some order of magnitude the solutions quality. Through the results analysis from the test cases made on the mutation and crossover prob-

abilities, it was found that the same configuration in all islands achieves better quality in the solutions. For the functions involved in the experiments, it was found that if all islands have the same diversification factors, the search leads to better quality of solutions.

This information is a preliminary experimental basis for other type of static and dynamic calibration experiments, in order to develop a self-adaptable environment for solving hard optimization problems.

References

1. Alba, E., Tomassini M.: Parallelism and Evolutionary Algorithms. In: Proc. of the IEEE Trans. on Evol. Comp., vol. 6, num. 5, pp. 443-462 (2002)
2. Ali, M., Pant, M., Nagar, A.: Two local Search Strategies for Differential Evolution. In: Conf. on Bio-Inspired Computing: Theories and Appl., pp 1429-1435 (2010)
3. Brest, J., Zamuda, A., Bokovie, B., Maucec M., Zumer, V.: High-Dimensional Real-Parameter Optimization using Self-Adaptive Differential Evolution Algorithm with Population Size Reduction. In: IEEE Congr. on Evol. Comp., pp 2032-2039 (2008)
4. Martínez, C., Rodriguez, F., Lozano, M.: Role differentiation and malleable mating for differential evolution: an analysis on large-scale optimisation. In: Soft Computing, vol. 15, issue 11, pp. 2109-2126 (2011)
5. Kozlov, K., Samsonov A.: New Migration Scheme for Parallel Differential Evolution. In: Proc. Int. Conf. on Bioinf. of Genome Reg. and Structure, pp 141-144 (2006)
6. Mattson T., Sanders B., Massingill B.: Patterns for Parallel Programming. Addison-Wesley, chapter 5, pp 143-152 (2004)
7. MPICH Message Passing Interface, http://www.mpich.org/
8. Noman, N., Iba, H.: Accelerating Differential Evolution Using an Adaptive Local Search. In: Proc. of the IEEE Trans. on Evol. Comp., vol. 12, pp 107-125 (2008)
9. Price, K., Stron R., Lampinen J.: Differential Evolution: A Practical Approach to Global Optimization. Springer. New York (2005)
10. Talbi, E.: Metaheuristics: From Design to Implementation. John Wiley & Sons, Hoboken, New Jersey (2009)
11. Tasoulis, D., Pavlidis, N., Plagianakos, V., Vrahatis, M.: Parallel Differential Evolution. In: Proc. of the Congr. Evol. Comp., vol. 2, pp. 2023-2029 (2004)
12. Zhao, S., Suganthan, P., Das, S.: Self-adaptive differential evolution with multi-trajectory search for large-scale optimization. In: Soft Computing - A Fusion of Foundations, Methodologies and App., vol. 15, num. 11, pp. 2175-2185 (2010)
13. Tang, K., Yao, X., Suganthan, P. N., MacNish, C., Chen, Y. P., Chen, C. M., Yang, Z.: Benchmark Functions for the CEC'2008 Special Session and Competition on Large Scale Global Optimization. Technical Report. In: Nature Inspired Computation and Applications Laboratory. USTC. China. pp. 4-31 (2007)
14. Yang, Z., Tang, K., Yao, X.: Self-adaptive Differential Evolution with Neighborhood Search. In: Proc. of the IEEE Congr. on Evol. Comp., pp. 1110-1116 (2008)
15. Yang, Z., Tang, K., Yao, X.: Scalability of generalized adaptive differential evolution for large-scale continuous optimization. In: Soft Computing, vol. 15, issue 11, pp. 2141-2155 (2011)
16. Zaharie, D., Petcu, D.: Adaptive Pareto Differential Evolution and Its Parallelization. In: Proc. of the 5th Int. Conf. Parallel Processing and Applied Mathematics, vol. 3019, pp. 261-268 (2004)

Cómputo en Paralelo para Integrales Multicéntricas usando una Distribución Balanceada

Ana Rosso[1], Claudia Denner[1], Guillermo Fraschetti[2], María Laura Tardivo[2,5], Jorge E. Pérez[3], Juan Cesco[4,5]

[1] Departamento de Matemática. Universidad Nacional de Río Cuarto.
(X5804BYA) Río Cuarto, Córdoba, Argentina
arosso@exa.unrc.edu.ar, cdenner@exa.unrc.edu.ar
[2] Departamento de Computación, Universidad Nacional de Río Cuarto.
(X5804BYA) Río Cuarto, Córdoba, Argentina
gfraschetti@dc.exa.unrc.edu.ar, lauratardivo@dc.exa.unrc.edu.ar
[3] Deparamento de Física. Universidad Nacional de Río Cuarto.
(X5804BYA) Río Cuarto, Córdoba, Argentina
eperez@exa.unrc.edu.ar
[4] Departamento de Matemática. Universidad Nacional de San Luis,
(D5700HHW) San Luis, Argentina.
[5] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
jcesco@unsl.edu.ar

Resumen La aproximación del campo autoconsistente (SCF) es utilizada en Química Computacional para abordar problemas de Química Cuántica. Entre otros, la energía molecular y la geometría de equilibrio son elementos determinados al nivel SCF. El enfoque SCF-LCAO presenta complejidad computacional, pues requiere gran esfuerzo de cálculo. Hemos desarrollado un programa secuencial, que posee características interesantes desde el punto de vista numérico, pero con alto costo en tiempo computacional. Con el objetivo de reducir el mismo se propuso paralelizar el algoritmo; actualmente se cuenta con una primera propuesta paralela. En este trabajo se presenta una nueva versión paralela del algoritmo, utilizando el modelo de comunicación y programación basado en pasaje de mensajes, empleando una técnica de descomposición de datos que contempla una carga de trabajo ponderada según el costo computacional de cada función. Los resultados obtenidos con esta nueva versión resultan satisfactorios en cuanto a la reducción del tiempo de cómputo.

Keywords: Cálculo Molecular, Aproximación SCF-LCAO, Paralelización, Balance de Carga.

1. El Tema en Estudio

Para conocer la estructura y las propiedades de una molécula es necesario estudiar la solución de la ecuación de Schrödinger, independiente del tiempo:

$$H\Psi = E\Psi \tag{1}$$

donde H es el operador Hamiltoniano, Ψ es la función de onda que describe el estado del sistema y E es la energía molecular. Se está interesado en encontrar soluciones aproximadas para la ecuación de autovalores (1), ya que la misma no puede resolverse exactamente, salvo en casos simples [20]. Uno de los procedimientos que se utiliza para obtener soluciones aproximadas está basado en la teoría de Hartree-Fock-Roothaan. A este procedimiento por ser iterativo se lo denomina "campo autoconsistente" (Self Consistent Field), y por provenir de combinaciones lineales de orbitales atómicos (Linear Combination of Atomic Orbitals), se lo denota por las siglas en inglés SCF-LCAO.

La complejidad computacional del método se centra en el gran número de integrales bielectrónicas que se deben calcular, aún en modelos de relativamente bajo tamaño. Este número es del orden de n^4, siendo n el número de orbitales que modelan la molécula.

Debido a su popularidad, SCF ha sido constantemente estudiado y optimizado para conseguir mejor desempeño computacional y hacer posible su uso en el estudio de sistemas de mayor tamaño. Los estudios realizados sobre el método SCF cubren, entre otros, cuatro aspectos importantes:

- Selección de una base de orbitales atómicos adecuada. Bases que han sido extensivamente estudiadas son las que incluyen orbitales atómicos tipo Slater (STO) [1], [2], orbitales Gaussianos (GTO)[3] y B-funciones [4], [5]. Una variante de los GTO la constituyen los orbitales Gaussianos flotantes (FGTO) [6],[7].
- Desarrollo de los cálculos relativos a la base introducida. Es un hecho ampliamente conocido que, cualquiera sea la base seleccionada, las integrales bielectrónicas resultantes, en la mayoría de los casos, no se pueden resolver en forma analítica. Por lo tanto, es necesario calcularlas numéricamente por métodos aproximados [8], [9], [10],[11], [12].
- Implementación de programas computacionales tendientes a obtener información sobre diferentes propiedades moleculares. En los últimos años se han desarrollado varios programas, algunos comerciales, otros desarrollados en centros científicos, los cuales realizan cálculos moleculares de distintos tipos. Sus implementaciones pueden ser seriales o en paralelo y sus capacidades de cálculo son diversas. *Gaussian* [25] es uno de los más difundidos. Otras alternativas se encuentran en [13], [14], [15].
- Aplicación de los programas para el estudio de sistemas concretos. Esta es un área muy activa de trabajo. Un punto importante en este campo es el desarrollo de heurísticas que sirvan de guía para la ubicación de los orbitales atómicos, tendientes a reproducir adecuadamente comportamientos que se conocen desde un punto de vista químico. No menos importante es el desarrollo de estrategias que permitan combinar estructuras simples resueltas para modelar sistemas más complejos [16], [17], [18], [19].

2. El Algoritmo de Cálculo SCF-LCAO (rasgos generales)

El algoritmo de cálculo que computa la aproximación SCF-LCAO permite utilizar distintos modelos de bases atómicas. A continuación se presenta de manera general, el esquema de procesamiento del algoritmo (para obtener un detalle completo del algoritmo y cada una de sus partes, ver [20]):

1. El programa requiere como dato de entrada una geometría inicial de la molécula a optimizar (coordenadas nucleares y el número de electrones).
2. Se calculan todas las integrales moleculares requeridas: $S_{\mu\nu}$, $H_{\mu\nu}$, $(\mu\nu|\sigma\lambda)$.
3. Diagonalizar la matriz de overlap $\mathbf{S}$ y obtener la matriz de transformación $\mathbf{X}$.
4. Obtener la matriz de densidad $\mathbf{P}$.
5. Calcular la matriz $\mathbf{G}$, a partir de la matriz de densidad $\mathbf{P}$ y de las integrales bielectrónicas $(\mu\nu|\sigma\lambda)$.
6. Obtener $\mathbf{F} = \mathbf{H} + \mathbf{G}$
7. Calcular $\mathbf{F}' = \mathbf{X}^*\mathbf{F}\mathbf{X}$
8. Diagonalizar $\mathbf{F}'$, para obtener $\mathbf{C}'$ y ϵ, y calcular $\mathbf{C} = \mathbf{X}\mathbf{C}'$
9. Cuando se comprueba la convergencia, el proceso finaliza, caso contrario vuelve a 4.

Con este algoritmo se obtiene la geometría molecular óptima minimizando la energía. Para ello, se realiza el cálculo anterior en diferentes posiciones relativas de los átomos.

3. Desarrollo de la investigación

El grupo de investigación desarrolló en una primera etapa una propuesta secuencial del algoritmo mencionado, el cual logra la optimización de geometrías moleculares a través de la minimización de la energía. Uno de los problemas de esa versión algorítmica es el alto tiempo de cómputo insumido, razón por la cual se han intentado diferentes alternativas a fin de mejorar esta implementación.

Una línea de trabajo estudia, fundamentalmente, variantes en las técnicas de cálculo para el algoritmo que modela el problema [16],[17],[18],[19],[26], [27], [28]. Otra se enfoca en paralelizar el algoritmo secuencial. Se trabajó inicialmente en paralelizar una regla de cuadratura gaussiana [30]. Posteriormente, se paralelizó una parte del algoritmo secuencial en el cual se distribuyen los cálculos intermedios en diferentes unidades de procesamiento, con una distribución de carga fija [29]; obteniéndose mejoras respecto de los tiempos secuenciales y un buen balance de carga para el caso de prueba utilizado.

En este trabajo, se presenta una modificación de la última versión paralela, respecto del mecanismo de asignación de carga a cada unidad de procesamiento. Aquí, la distribución se realiza en función de un peso asociado al tipo de cálculo a realizar. A diferencia de la versión anterior, esta nueva propuesta considera las particularidades de la molécula. Con este esquema, se posee mayor control del costo computacional en relación al ejemplo en estudio.

A continuación daremos una descripción de la base de funciones utilizadas en la implementación, y posteriormente presentaremos la nueva versión paralela.

4. Funciones utilizadas en la modelización

Una buena base de orbitales atómicos debe satisfacer dos condiciones para dar soluciones apropiadas. Dichas condiciones vienen dadas por el "comportamiento cuspidal alrededor de los núcleos" y el "decaimiento exponencial" en el infinito [21].

En el desarrollo del programa secuencial se utilizó una base de funciones, con orbitales atómicos 1s de Slater (STO) y gaussianos (GTO). Dichos orbitales tienen las siguientes expresiones:

$$\Phi_\mu(\vec{r}) = C' e^{(-\alpha_\mu |\vec{r} - \vec{R_\mu}|)} \qquad \text{(STO)}$$

$$\Phi_\mu(\vec{r}) = C e^{(-\alpha_\mu |\vec{r} - \vec{R_\mu}|^2)} \qquad \text{(GTO)}$$

donde $C = (\frac{2\alpha_\mu}{\pi})^{3/4}$, $C' = (\frac{\alpha_\mu^3}{\pi})^{1/2}$, $\vec{r}$ pertenece a $\mathbb{R}^3$, el vector $\vec{R_\mu}$ se denomina centro del orbital y α_μ es el coeficiente orbital.

Las funciones GTO antes mencionadas, a pesar de no satisfacer la condición de comportamiento cuspidal alrededor de los núcleos, son las más utilizadas, puesto que las integraciones necesarias se realizan eficientemente. En contraste, las funciones STO satisfacen ambas condiciones, pero los cálculos de las integraciones son computacionalmente costosas.

En la implementación secuencial realizada, la mayor parte del tiempo de cálculo insumido lo requiere la evaluación del potencial promedio de Hartree-Fock, pues cuando se introduce la base atómica, hay que evaluar integrales bielectrónicas multicéntricas que tienen la siguiente expresión:

$$(\mu\nu|\sigma\lambda) = \int\int_{\mathbb{R}^3 \times \mathbb{R}^3} \Phi_\mu(\vec{r_1})\Phi_\nu(\vec{r_1}) \frac{1}{|\vec{r_1} - \vec{r_2}|} \Phi_\sigma(\vec{r_2})\Phi_\lambda(\vec{r_2}) d\vec{r_1}\, d\vec{r_2} \qquad (2)$$

donde los índices $a = \mu, \nu, \sigma$ y λ corresponden a las funciones $\Phi_a(\vec{r_i})$ de la base dadas por los orbitales (STO) y (GTO). A la expresión $(\mu\nu|\sigma\lambda)$ la llamamos *cantidad de cuatro índices*.

Esta particularidad de utilizar una base formada por dos clases de funciones es lo que diferencia nuestra implementación de otras, como por ejemplo las que se utilizan en el programa de cómputo *Gaussian* [25].

En nuestra implementación, los orbitales STO se posicionan en los núcleos, para modelar el correcto comportamiento de la función de onda. Los orbitales GTO pueden ser colocados en los núcleos o en los enlaces que unen dos núcleos. Los resultados obtenidos mediante esta modelización han sido competitivos en cuanto a la geometría y a las cifras significativas.

En estos cómputos cuando $\Phi_\mu, \Phi_\nu, \Phi_\sigma$ y Φ_λ son de tipo GTO, la integral de la fórmula (2) posee una primitiva conocida, por lo que su cálculo es directo. En aquellos casos en los que la expresión (2) utiliza al menos una función del tipo STO, el cómputo requiere la aproximación numérica de las integrales, pues ellas no tienen expresión analítica [9]. Estas integrales pueden ser unidimensionales, bidimensionales o tridimensionales.

5. Modelo de Cómputo en Paralelo

5.1. El Programa

En nuestro programa secuencial, el cómputo de las integrales de cuatro índices de la fórmula (2) guía su procesamiento a través de un vector, en el cual se disponen todas las combinaciones de funciones a evaluar. Estas combinaciones de funciones son independientes unas de otras, por lo cual es posible aplicar un esquema de paralelización directo.

La primera versión de la implementación en paralelo consideraba, para la distribución de carga, sólo la cantidad de funciones involucradas en el cálculo, de manera de asignar cierto porcentaje de esas funciones a cada procesador. En nuestro problema las funciones no poseen la misma complejidad computacional, por lo cual, el tiempo que requiere su procesamiento es variable. Para compensar esta variabilidad, en esta primera versión paralela se consideraron porcentajes no equitativos en la distribución de carga. En una batería de pruebas se utilizaron diferentes valores para estos porcentajes de manera tal de analizar el comportamiento del algoritmo con este esquema de paralelización ad-hoc. Con esta distribución se obtuvieron resultados aceptables en cuanto a la reducción del tiempo de cómputo, manteniendo la calidad de la aproximación [29]. Sin embargo, esta distribución de carga resulta ser más adecuada cuando el costo computacional de cada una de las funciones es equitativo.

La nueva versión del programa paralelo, al igual que en la versión anterior, sigue un modelo algorítmico *master-worker* [22], en donde el proceso *master* es el encargado de inicializar los cálculos, generar las matrices iniciales y preparar los datos para distribuir el cómputo entre los procesadores *workers*. Luego de finalizar el procesamiento que les fue asignado, los *workers* retornan al proceso *master* sus resultados parciales, para que éste útlimo obtenga la solución final.

En esta segunda versión se realiza la distribución de carga clasificando cada función según el costo computacional del trabajo matemático involucrado, parámetro que llamamos *peso*. A partir de esto se distribuyen los cálculos utilizando como criterio el balance de la suma de los pesos, en lugar de considerar solamente la cantidad de funciones.

Para la ponderación, el cálculo de coeficientes $(\mu\mu|\nu\nu)$ de las cantidades de cuatro índices que involucra solo funciones de tipo STO (en centros distintos o centros coincidentes), y todas las cantidades de cuatro índices donde solo intervienen funciones GTO, requieren operaciones de suma, multiplicación y evaluación de funciones simples, siendo estos los coeficientes de menor costo computacional. A los mismos les asignamos el menor peso.

El cálculo de coeficientes $(\mu\nu|\sigma\lambda)$, con todas funciones de tipo STO, donde todos los índices son distintos, involucra además de las operaciones elementales, la evaluación de una integral triple, que es aproximada usando la regla de cuadratura de Gauss-Legendre con 32 puntos en cada variable. A estos cálculos, que requieren el mayor costo computacional, le asignamos el mayor peso.

Como el *peso* de cada función dependen del tipo de cálculo a realizar y no de una molécula y/o función en particular, los mismos pueden establecerse en forma previa a la ejecución del programa.

Si bien se conoce la complejidad computacional de cada función, la cantidad y el tipo de funciones de la base que modelan una molécula varía de acuerdo al compuesto químico que se esté abordando. Por este motivo, se intenta obtener un esquema de mapeo dinámico, en el sentido de establecer que el proceso master determine, considerando el peso y la cantidad de funciones, cuál/les son las tareas asignadas a cada worker.

La vesión paralela utiliza el lenguaje Octave [24], software libre de cálculos numéricos. Se decidió utilizar un modelo de comunicación y programación basado en pasaje de mensajes, bajo una organización de memoria distribuida [22]. Para su implementación se empleó la librería MPI [23], en conjunto con el lenguaje de programación C. Se utilizaron paquetes adicionales para invocar, desde el código escrito en el lenguaje C, rutinas escritas en Octave (mex-interface) [24].

5.2. Los Ensayos y Resultados

A continuación se muestra un ejemplo de una molécula donde se fijan sobre los núcleos las funciones STO y sólo se varía la cantidad de funciones GTO. Las funciones STO utilizadas en las pruebas se centraron en los siguientes núcleos:

$\overrightarrow{R_\mu} = (0, 1{,}4296, 1{,}1124)$,
$\overrightarrow{R_\nu} = (0, -1{,}4296, 1{,}1124)$,
$\overrightarrow{R_\sigma} = (0, 0, 0)$ con coeficientes orbitales $\alpha_\mu = 1, \alpha_\nu = 1, \alpha_\sigma = 8$

Modelo	*Total de Funciones*	*Tiempo Secuencial*	*Tiempo Paralelo*
1	15 *(3 STO, 12 GTO)*	0m 9.07s	0m 7.85s
2	25 *(3 STO, 22 GTO)*	3m 55.62s	1m 41.68s
3	50 *(3 STO, 47 GTO)*	750m 20.34s	350m 49.79s

Cuadro 1. Tiempos de ejecución con las dos versiones algorítmicas.

El cuadro 1 muestra las aceleraciones obtenidas con la máxima capacidad de cálculo disponible. Se contrastan en la misma los tiempos obtenidos con el programa secuencial, incluyendo tres modelos distintos de la misma molécula. Se

puede observar, en todas las ejecuciones, que la versión paralela obtiene menores tiempos de cómputo.

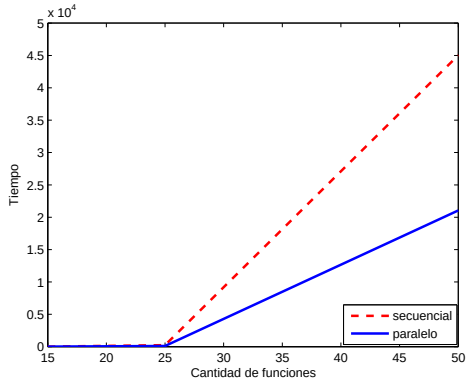


Figura 1. Tiempo de ejecución (en segundo) vs cantidad de funciones de la molécula. Se puede observar que el modelo secuencial insume más tiempo de cómputo que el modelo paralelo. La tendencia indicaría que al aumentar la cantidad de funciones que modelan la molécula, el modelo paralelo sigue teniendo mejor desempeño.

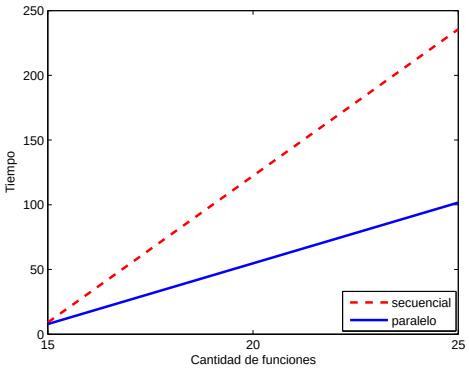


Figura 2. Tiempo de ejecución (en segundos) vs cantidad de funciones de la molécula. En esta figura se muestra en detalle el primer intervalo de la Figura 1, donde se puede constatar un comportamiento similiar al descripto anteriormente para moléculas que utilizan entre 15 y 25 funciones en su modelado.

Las Figuras 1 y 2 grafican el comportamiento de los valores obtenidos en la tabla 1.

Las pruebas piloto fueron llevadas a cabo utilizando una computadora con procesador Intel(R) Core(TM) i7-2600, con 8 procesos workers más un proceso adicional asignado al master. Si bien el código está preparado para realizar

pruebas en un entorno real de memoria distribuída, como por ejemplo un cluster de computadores, al momento disponemos solamente del recurso mencionado.

6. Reflexiones Finales

En este trabajo se presenta una nueva versión paralela de un programa secuencial dedicado al cálculo molecular, para determinar la función de onda y la mínima energía en la aproximación SCF-LCAO. Esta versión considera, para la distribución del trabajo, la suma de los pesos relativos de cada función involucrada en el cálculo. Con ella se obtuvieron buenos resultados en cuanto a reducción del tiempo de cómputo, manteniendo la calidad de la aproximación.

Las pruebas realizadas aún no fueron escaladas a mayor cantidad de procesadores, por ejemplo, a través de la utilización de clusters. Como trabajo futuro se planea la ejecución del código desarrollado en un entorno con mayor cantidad de unidades de cómputo. A la vez se espera desarrollar una nueva versión paralela, considerando un modelo de asignación de tareas y balance de carga dinámico, en el cual los procesos worker tomen los trabajos a partir de un *pool* de tareas. Con este nuevo enfoque se intentará desarrollar un modelo que continue considerando la carga de trabajo variable en cada unidad de cálculo.

Agradecimientos

A la Secretaría de Ciencia y Técnica de la Universidad Nacional de Río Cuarto, y al Ministerio de Ciencia y Tecnología de la Provincia de Córdoba, por los recursos puestos a nuestra disposición para llevar adelante esta investigación.

Referencias

1. Slater, J.C.: Atomic Shielding Constants, Physical Review, vol. 36, issue 1, pp. 57-64 (1930)
2. Roothaan, C.: Reviews of Modern Physics, vol. 23, issue 2, pp. 69-89 (1951)
3. Boys, S.:Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. In: Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, vol. 200, pp.542-554 (1950)
4. Steinborn, E., Weniger, J.: Advantage of reduced bessel functions as atomic orbitals. An application to H+2. International Journal of Quantum Chemistry. In: Quantum Chemistry Symposium, vol. 11, pp. 509 (1977)
5. Steinborn, E.,Weniger, J.: Reduced bessel functions as atomic orbitals: Some mathematical aspects and an LCAO-MO treatment of HeH^{++}. In: International Journal of Quantum Chemistry: Quantum Chemistry Symposium, vol. 12, pp. 103-108 (1978)
6. Maggiora, G., Christoffersen, J.: Ab initio calculations on large molecules using molecular fragments. Generalization and characteristics of floating spherical Gaussian basis sets. In: Journal of the American Chemical Society, vol. 98, issue 26. pp. 8325-8332 (1976)

7. Pérez, J., Cuenya, H., Contreras, R. Ortiz, F., Grinberg, H., Ruiz de Azua, M., Giribet, C.: Expansion of atomic orbital products in terms of a complete function set. In: Theoretica Chimica Acta A Journal for Structure, Dynamics and Radiation, vol.88, number 2, pp. 147-168 (1994)
8. Bouferguene, A., Fares, M., Rinaldi, D.: Integrals over B functions basis sets. I. Three-center molecular integrals, a numerical study. In: Journal of Chemical Physics, vol. 100, pp. 8156-8119 (1994)
9. Shavitt, I., Karplus, M.: Gaussian-Transform Method for Molecular Integrals. I. Formulation for Energy Integrals. In: Journal of Chemical Physics, vol 43, pp. 398-415 (1965)
10. Montagnani, R., Salvetti, O.: Computation of many-center exchange integrals over Slater orbitals up to 4d by means of optimized Gaussian expansions. In: International Journal of Quantum Chemistry, vol. 47, issue 3, pp.225-229 (1993)
11. Fernandez Rico, J., Lopez, R., Ramírez, G., Tablero, C.: Molecular integrals with Slater basis. V. Recurrence algorithm for the exchange integrals. In: Journal of Chemical Physics, vol. 101, pp. 9807-9817 (1994)
12. Safouhi, H., Hoggan, P.: Recent progress in the accurate and rapid evaluation of all Coulomb integrals over slater-type orbitals .In: International Journal of Quantum Chemistry , vol. 84, pp. 580-591 (2001)
13. Fernandez Rico, J., Lopez R., Ema, I., Ramírez, G.: Calculation of many-centre two-electron molecular integrals with STO. In: Computer Physics Communications, vol. 105, pp. 216-224 (1997)
14. Bouferguene, A., Fares, M., Hogan, P.: STOP: A slater-type orbital package for molecular electronic structure determination. In: International Journal of Quantum Chemistry, vol. 57, pp. 801-810 (1996)
15. Mantovani, M., Malagoli, M.: Highly Parallel SCF Calculation: the SYSMO Program. In: Proceedings of the Parallel and Distributed Processing, CICAIA, Modena University, Italia, pp. 502-507 (1995)
16. Pérez, J., Cesco, J., Taurian, O., Ortiz, F., Rosso, A., Denner, C., Giubergia, G.: A new algorithm to evaluate bielectronic integrals with 1s Slater type orbitals obtained by the integral transforms. In: International Journal of Quantum Chemistry, vol 99, pp.70-79, (2004)
17. Cesco, J., Pérez, J. , Denner, C., Giubergia, G., Rosso, A.: Rational approximants to evaluate four-center electron repulsion integrals for 1s hydrogen Slater type functions. In: Applied Numerical Mathematics, vol 55, pp. 173-190 (2005)
18. Pérez, J., Cesco, J., Taurian, O., Ortiz, F., Rosso, A., Denner, C., Giubergia, G.: Evaluation of Bielectronic Integrals 1s Slater Orbitals by using Averages. In: International Journal of Quantum Chemistry, vol 102, pp. 1056-1060 (2005)
19. Pérez, J., Taurian, O., Cesco, J., Ortiz, F.: A New Method for approximating the Coulomb Potential Generated by Product of two 1s Slater Orbitals. In: Quantum Chemistry Research Trends, pp 215-227, (2007)
20. Szabo,A., Ostlund, N.: Modern quantum chemistry. Introduction to advanced Electronic Structure Theory, Macmillan Publishing Co, Inc., New York (1982)
21. Kato, T.: On the eigenfunction of many-particle system in quantum mechanics. In: Communication on Pure and Applied Mathematics, vol. 10, issue 2, pp. 151-177 (1957)
22. Grama, A., Gupta, A., Karypis, G., Kumar, V.: Intruduction to Parallel Computing, Second Edition. Pearson Addison Wesley (2003)
23. Message Passing Interface, http://www.open-mpi.org
24. http://www.gnu.org/software/octave

25. Gaussian Software, http://www.gaussian.com/
26. Denner, C.: Una aplicacion de aceleradores de convergencia al calculo de integrales multicentricas en la teoria de Hartree-Fock. Biblioteca Juan Filloy. Universidad Nacional de Río Cuarto (2004).
27. Rosso, A.: Técnicas de desacople en el cálculo de integrales multicéntricas. Universidad Nacional de Río Cuarto (2004).
28. Giubergia. G.: Metodos alternativos para el calculo de integrales multicentricas de tres y cuatro centros en la teoria de Hartree - Fock. Universidad Nacional de Río Cuarto (2001).
29. Rosso, A., Denner, C., Fraschetti, G., Tardivo, L., Pérez, J., Cesco, J.: Paralelización del cálculo molecular de las integrales bielectrónicas en la aproximación SCF-LCAO. In: IV Congreso de Matemática Aplicada, Computacional e Industrial (MACI) (2013)
30. Rosso, A., Denner, C., Daniele, M., Fraschetti, G.: Implementación con MPI de reglas adaptivas de cuadratura. In: III European-Latin-American Workshop on Engineering Systems. III SELASI. Universidad de Talca, Curicó, Chile (2007)

Predicting the communication pattern evolution for scalability analysis

Javier Panadero*, Alvaro Wong, Dolores Rexachs, and Emilio Luque

Computer Architecture and Operating System Department, Universitat Autonoma of Barcelona, Barcelona, SPAIN.
{javier.panadero, alvaro.wong}@caos.uab.es,
{dolores.rexachs, emilio.luque}@uab.es

Abstract. The performance of the message-passing applications on a parallel system can vary and cause inefficiencies as the applications grow. With the aim of providing scalability behavior information of these applications on a specific system, we propose a methodology that allows to analyze and predict the application behavior in a bounded time and using a limited number of resources. The proposed methodology is based on the fact that most scientific applications have been developed using specific communicational and computational patterns, which have certain behavior rules. As the number of application processes increases, these patterns change their behavior following specific rules, being functionally constants. Our methodology is focused on characterizing these patterns to find its general behavior rules, in order to build a logical application trace to predict its performance. The methodology uses the PAS2P tool to obtain the application behavior information, that allow us to analyze quickly a set of relevant phases covering approximately 95% of the total application. In this paper, we present the entire methodology while the experimental validation, that has been validated for the NAS benchmarks, is focused on characterizing the communication pattern for each phase and to model its general behavior rules to predict the pattern as the number of processes increases.

Keywords: Prediction Scalability, Communication Pattern, MPI applications

1 Introduction

During the last years, due to constant hardware evolution, high performance computers have increased the number of cores significantly. Users of these systems want to get the maximum benefit of the number of cores and scale their applications, either by reducing the execution time or increasing the workload.

* This research has been supported by the MICINN Spain under contract TIN2007-64974, the MINECO (MICINN) Spain under contract TIN2011-24384, the European ITEA2 project H4H, No 09011 and the Avanza Competitividad I+D+I program under contract TSI-020400-2010-120.

To achieve an efficient use of high performance systems, it would be important to consider the analysis of the application behavior, before executing an application in a large system, since the ideal number of processes and resources required to run the application may vary from one system to another, due to hardware differences. The lack of this information may cause an inefficient use, causing problems at different levels, such as not achieving the expected speed up, or increasing the economic and energy cost. To avoid these problems and make an efficient system use, users and system administrators use predictive performance models selecting the most appropriate resources to run the application.

In this paper, we propose a methodology that will allow us to analyze and predict the scalability behavior for message-passing applications on a given system, in a bounded time and using a reduced set of resources. The objective of the methodology is to predict the application performance when increasing the number of processes, characterizing and analyzing the behavior of the communication and computation patterns.

The methodology is based on the fact that most scientific applications have been developed using specific patterns, which have functional similarity according to the number of processes, following behavior rules. To characterize these rules, we used the PAS2P methodology [1]. PAS2P identifies the application phases, which contain a specific communication pattern and allows us to reduce the complexity of the application analysis by creating the application signature, which contains the relevant communication and computation patterns of the application (phases), and their repetition rates (weights).

The application phases can be observed and analyzed during the execution time dynamically, this is without having the source codes, to relate them by functional similarity increasing the application processes, with the objective to model their general behavior rules, to build an application logical trace that will be independent of the machine. Once we have the logical trace, the last step is to convert this trace in machine dependent, through the machine parameters to get the physical trace This new trace will be used to predict the performance of the application.

This paper is organized as follows: Section II presents the related work, Section III presents an overview of the PAS2P methodology, Section IV presents the proposed methodology, Section V presents the experimental validation, which is focused in the communication pattern modeling, and finally Section VI presents the conclusions and future work.

2 Related Work

There are other works, related to the study of communication patterns of MPI applications. I. Lee et al [3] proposes to analyze the communication patterns of NAS-MPI benchmarks to understand the communication behavior in scientific workloads and to predict the larger scale program behavior. This work is focused on measuring the communication timing, the sources and destinations and mes-

sage sizes. This work differs from our proposal in that we model the general rules of the communication pattern and communication volume.

R. Preissl et al [4] presents an algorithm for extracting communication patterns. The algorithm finds locally repeated sequences on each node using a suffix tree algorithm and matches these local repetitions with other sequences on other nodes to generate a global pattern. This approach differs of our work in that we use a functional similarity algorithm, instead of the suffix tree algorithm.

M. Chao et al [5] proposes a methodology to determine the communication pattern similarities between two programs using two metrics to form a coordinate on a 2-dimensional Cartesian Space. Our work differs of this, because we search the similarity as the number of processes increases.

3 Overview about PAS2P methodology

The PAS2P methodology [1] is composed by two steps. The first step is done on a base machine and consists on analyzing the application, building the application model to extract its phases and weights that will use to construct the signature, which is an executable that contains the application phases. The second step consists on executing the signature on a target system, to measure the execution time of each phase. Once these times have been measured, equation 1 is used to predict the application execution time in the target system, where PET is the Predicted Execution Time, n is the number of phases, TEPhasei is the Phase i Execution Time and Wi is the weight of the phase i.

$$PET = \sum_{i=1}^{n} (TEPhasei)(Wi) \tag{1}$$

4 Proposed Methodology

Fig. 1 shows the proposed methodology, which is based on the fact that most message-passing applications have been developed using specific communication and computation patterns, which have functional similarity when the number of processes increases, following specific rules of behavior. These patterns compose the application phases, that can be observed and traced to relate them when the number of processes increases, in order to find and model their general behavior rules. Once the patterns have been modeled, it is possible to generate the logical trace, which will provide the communication and computation times to predict the application performance.

In this section, we will describe each of the methodology stages, focusing on the characterization and modeling of the communication pattern, which is the objective of this paper.

4.1 Characterization

The characterization step comprises two sub-stages: Machine Characterization and Application Characterization. The Machine Characterization is done once, regardless of the application which scalability we will predict.

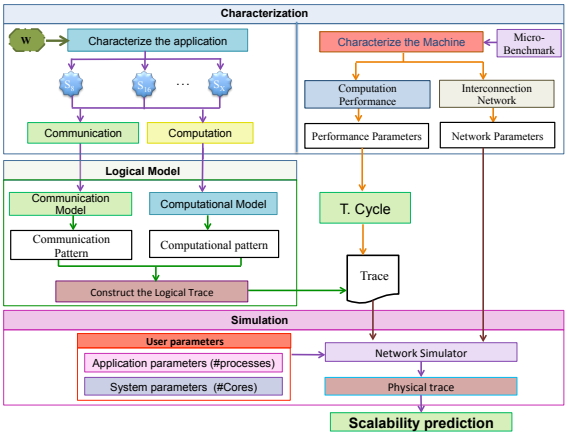


Fig. 1: Proposed Methodology

Machine Characterization. Consists on characterizing the machine performance in computation and communication. First of all, the performance is analyzed to obtain the real cycle time, when the processor is executing floating-point instructions and there are not misses on any level of the memory hierarchy. We characterize instead of using the theoretical time provided by the manufacturer, because there may be differences between them. In order to obtain the real cycle, a micro-benchmark was developed. On the other hand, the interconnection network is characterized using benchmarks.

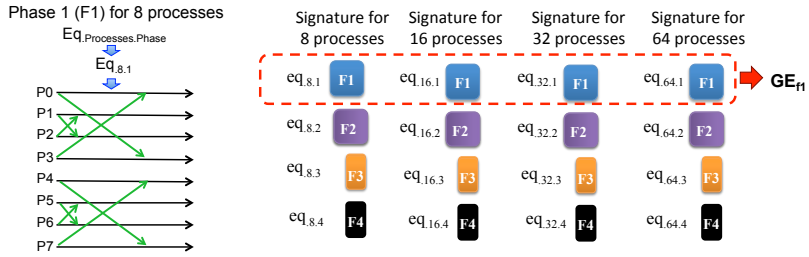
Application characterization. Consists on analyzing the application behavior (communication and computation) to obtain information and build its logical trace. We carry out a set of signature executions for a small and different number of processes, that will be analyzed to extract information of each phase of the signature. Each phase identifies a repetitive computation and communication behavior. The application signatures are obtained with PAS2P tool [2], which is a tool that applies the PAS2P methodology in an automatic and transparent way. It was decided to work with the signature rather than the whole application, since the signature contains only the relevant application phases. By analyzing this small set of relevant phases, we cover about 95% of the whole application. For this work, we integrated PAS2P with PAPI hardware performance tool [6] to obtain low-level performance information, such as the number of instructions and the number of cycles of each phase.

4.2 Logical trace generation

Once the relevant phases have been characterized, the communication and computation patterns are modeled for each phase to obtain the general behavior rules, which will be used to generate the application logical trace. This trace will be machine independent, according to how the application has been developed.

The parameters of the general behavioral rules will define the trace for a specific number of processes. Once the logical trace has been generated, the predicted computation and communication times are provided to generate the physical application trace, which will be dependent on the machine and will allow us to predict the application performance.

Communication pattern modeling. The communication pattern comprises the general behavior equations, which predict the destination from the source, and the data volume for each phase. The phases will be related by functional similarity between the signatures with different number of processes, in order to model their behaviors. It should be mentioned that a phase can have 1 to N communications, depending on the application. The predicted data volume of each communication will be obtained by mathematical regression models, while for obtaining the general communication rules (source - destination), we propose an algorithm, based on obtaining the communication equations that calculate the destination from the source ($\text{eq}._{processes.phase}$) for each phase of the signatures (local equations). From these equations, using functional similarity, the general equation behavior is modeled. We show an example of this algorithm, which considers that each phase (F_i) has only one communication, and therefore one local equation, as shown in fig. 2(a) for the phase 1 for 8 processes. Once the local equations have been obtained for each phase of each signature, these are analyzed by functional similarity to model the general behavior equation (GE_{Fi}) as shown in fig. 2(b) for phase 1.



(a) F1 for 8 processes (b) Comparison of phases by functional similarity

Fig. 2: Obtaining general equations by functional similarity

The algorithm to obtain the behavior rules is divided in two stages, in the first stage the local equations are generated for each phase, and in the second stage the general equations are obtained, as shown in fig. 3. Noteworthy that for both steps, the process identifier (process number) is converted to binary in order to work at bit level.

The first algorithm stage is composed of a first sub-stage of analysis and a second sub-stage of modeling. During the analysis phase, the dependencies between processes (Dependent, No-Dependent), the pattern type: Static (Mesh,

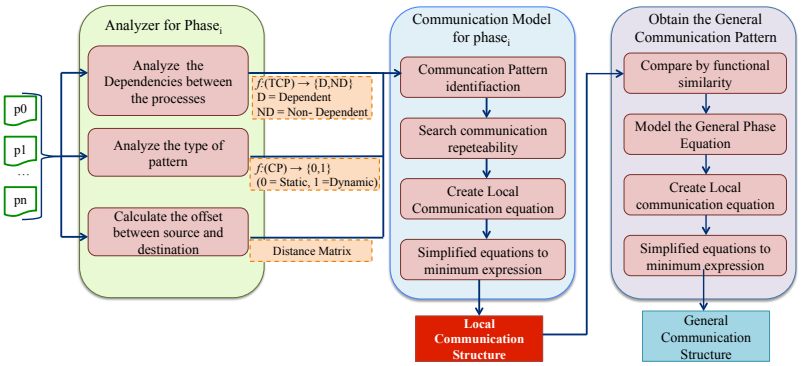


Fig. 3: Proposed algorithm to model the communication pattern

Ring, ...) or Dynamic (Exchange or Permutation), and the distance matrix between processes are obtained for each phase of the signatures.

All this information is provided to the modeling phase to generate the local equations. In this second sub-stage, the communication pattern of each phase is identified. Moreover, the repeatability of the communication is sought to generate a more structured equation model. Once this information has been identified, the local equation for each communication is generated and applied in a compression method to simplify the next step of modeling. The output of this module is a communication structure that identifies each local equation. The algorithm uses two different structures because the way to predict the communication pattern is different depending on the pattern type. If the pattern is dynamic, the obtention of the destination process is based on the exchange of a certain number of source bits, which are called bits involved. For this type of algorithm, the first structure (EC1) is used. In case of a static pattern, to obtain the destination process we search the repeatability of the communications, for example, in a 4 x 4 mesh, the first three processes in the first row have a displacement of 1, while the fourth process connects with the first and has a displacement of 3. This behavior is repeated for the remaining rows of the mesh. For this type of patterns, the second structure (EC2) is used.

The structure EC1 has the number of phase (#Phase), the communication number of the phase (#Comm), the algorithm type (Exchange, Permutation) and a vector with the bits involved in the pattern as parameters, while the structure EC2 has the number of phase, the communication number of the phase, and a list of communication and number of repetitions as parameters.

1. EC1 = {#Phase, #Comm, Algorithm Type , List of bits involved }

2. EC2 = {#Phase, #Comm, list[communication list[#repetition]] }

Fig. 4 shows a brief example of the procedure. We have a phase with 8 processes and three communications, where each communication has its own communication pattern. If we focus on the first communication, that shows its communication pattern, we generate the matrix distance between the source and

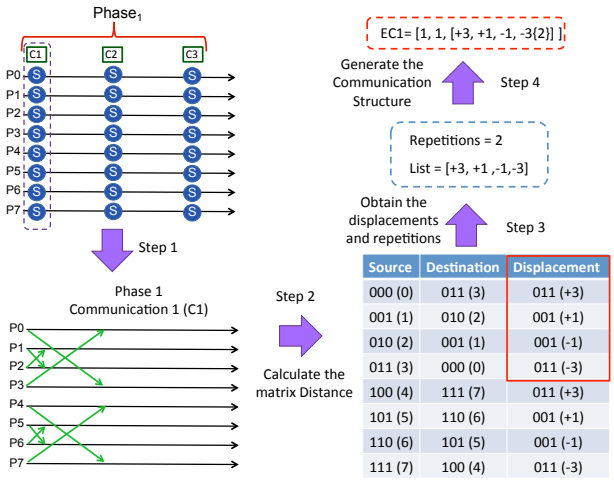


Fig. 4: Example of generating the Communication Structure

destination. The algorithm is static, then we search for repetitions, in this case, we have two, therefore the sequence {+3,+1,-1,-3} is repeated two times, once for processes from 0 to 3 and then for processes from 4 to 7. Once we have the sequences and repeatability, we create the local equation and generate the structure of communication EC1.

Once the communication structure has been obtained, it is submitted to the second algorithm stage with the purpose of obtaining the general equations, which model the general behavior of the communication pattern. To model the general equations of each phase, the local equations are analyzed by functional similarity. When the similarity has been identified, the general equations are modeled. The general equation has as variables the number of processes to predict, the displacements between the source and destination for the number of processes to predict, and the hops between the processes executed and the number of processes to predict. Finally, the last step that compresses the general equation in order to simplify the expression used to predict the communication pattern for a greater number of processes.

Computation pattern modeling. As shown in fig. 5, the algorithm is based on searching the repeatability of MPI primitives for each phase to identify the computation patterns. Then, they are compared using functional similarity between the different signature executions to predict the computation patterns for a larger number of processes. We search repeatability because these primitives are enclosed in repeated loops throughout the phase. The aim of the proposed algorithm, is to discover the minimum set of primitives and predict their repetition number to generate the logical trace, as it may vary depending on the number of processes. Once the information has been characterized, the computation time between the MPI primitives is modeled, based on separating the computation

time in: execution time, the time it takes the processor to execute instructions, the stall time and the time that the processor waits for memory accesses. We decided to separate computation times as they may have different trends when the number of processes increases. In order to separate these times, the model uses information from the hardware counter, obtained in the characterization stage. Then, the computation time will be predicted using statistical regression models .

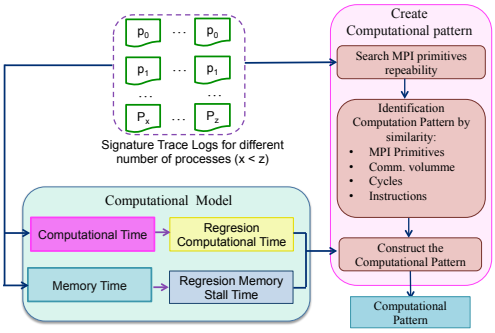


Fig. 5: Procedure to model the computation pattern

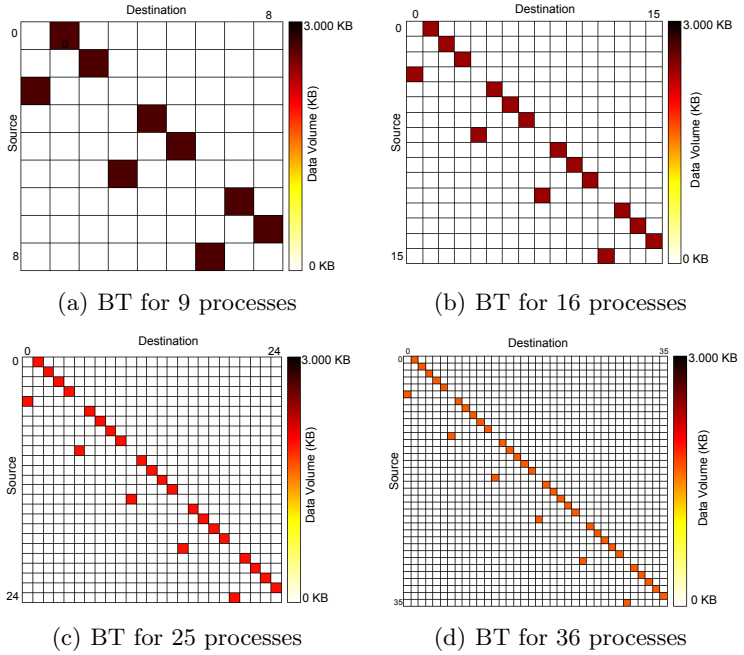
4.3 Simulation

Finally, in the last stage of the methodology we carried out a simulation of the trace in order to obtain the communication times of the messages from the trace on the physical machine. After the simulation, we obtained the runtime of each predicted phase that comprises the trace. Each phase time will be multiplied by the predicted weight of the phase to obtain the run time of the application for the number of processes we want to predict, as shown in Equation 1.

5 Experimental Validation

This section shows the experimental validation of the communication pattern modeling using the BT from the NPB class D. As experimental environment, we used a cluster of 16 nodes with 16 processors Intel Xeon quad-core.

To carry out the experimental validation, we executed a set of BT signatures and different number of processes (9, 16, 25 and 36). We predicted for 49 processes. The signatures were characterized to obtain the local equations for all phases and validated the proposed algorithm for all the communications in each phase, but due to a lack of space we only show the first communication in the first phase, as is shown in fig. 6. From this characterization, table 1 shows the local equations and the communication volumes. The pattern type is static (Toroidal), hence, the EC1 output structure is used. From the information of this structure, the general equation and the communication volume for each communication phase were modeled and validated for 49 processes.



(a) BT for 9 processes (b) BT for 16 processes

(c) BT for 25 processes (d) BT for 36 processes

Fig. 6: Characterization of the communication pattern and volume data

Table 1: Sumary of commnications for the different signatures

Signature (#Processes)	Phase	Com.	Local Equation	Communication Volume (KB)
9	1	1	[+1 {+2,+3}], [-2{+1,+3}]	2892
16	1	1	[+1 {+3,+4}], [-3{+1,+4}]	2496
25	1	1	[+1 {+4,+5}], [-4{+1,+5}]	2132
36	1	1	[+1 {+5,+6}], [-5{+1,+6}]	1849

Fig. 7 shows the general equation, corresponding to a static algorithm (Toroidal), which has a displacement +1 (first term of the equation), except for the processes located at the end, which connect with the initial nodes (second term of the equation). The variable *Disp* has a value of 1, since the distance between the source and destination is 1, while the variable K indicates the number of hops from the last characterized signature until the number of processes we want to predict. BT signature has been executed for 36 processes, and we want predict for 49 processes, then K has increased by 1 unit (1 hop). This is, by the application constraint, the next incremental step of 36 processes is 49, because BT only accepts processes of a square number as valid. On the other hand, in fig. 7 we show the communication volume equation, which is a potential regression and has a R-squared value of 0,97. If we apply this equation for 49 processes, we obtain a communication volume of 1703.59 KB. If we compare this value with the real execution, where we obtained a communication volume of 1628 KB, the prediction error is about 4.6%.

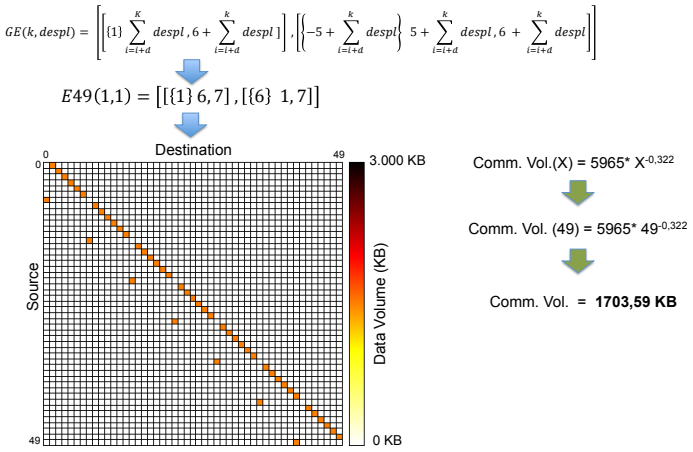


Fig. 7: Prediction of communication pattern and data volume for BT 49

6 Conclusions and future work

This paper proposes a methodology to analyze and predict the scalability behavior in message-passing applications in a given system, using a limited number of resources and bounded time. The methodology was presented and the modeling communication was experimentally validated. Currently, we are working on finishing the computational model validation and generating the logical trace, in order to insert it in a network simulator and obtain the physical trace, which will predict the application performance.

References

1. Wong, A., Rexachs, D., Luque, E.: Extraction of parallel application signatures for performance prediction. HPCC 10th (2010) 223–230
2. Panadero, J., Wong, A., Rexachs, D., Luque, E.: A tool for selecting the right target machine for parallel scientific applications. ICCS **18**(0) (2013) 1824 – 1833
3. Lee, I.: Characterizing communication patterns of nas-mpi benchmark programs. In: Southeastcon, 2009. SOUTHEASTCON '09. IEEE. (2009) 158–163
4. Preissl, R., Kockerbauer, T., Schulz, M., Kranzlmuller, D., Supinski, B., Quinlan, D.: Detecting patterns in mpi communication traces. In: Parallel Processing, 2008. ICPP '08. 37th International Conference on. (2008) 230–237
5. Ma, C., Teo, Y.M., March, V., Xiong, N., Pop, I., He, Y.X., See, S.: An approach for matching communication patterns in parallel applications. In: Parallel Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on. (2009) 1–12
6. Dongarra, J., Malony, A.D., Moore, S., Mucci, P., Shende, S.: Performance instrumentation and measurement for terascale systems. In: European Center for Parallelism of Barcelona. (2003) 53–62

Arquitectura en capas para acceso remoto SAD

Karina Cenci – Leonardo de - Matteis – Jorge Ardenghi

Laboratorio de Investigación en Sistemas Distribuidos
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
{kmc,ldm,jra}@cs.uns.edu.ar

Resumen La utilización de fuentes de datos compartidas facilita el trabajo en equipo. Por este motivo las organizaciones tienen implementan en sus locaciones sistemas con controles de acceso para compartir los datos de acuerdo a los privilegios de los usuarios. Nuevas formas de trabajo son la distribución de los miembros de un equipo en distintos lugares físicos, el trabajo desde las casas de los empleados, el traslado temporal a otra locación. Por todas estas razones, acceder a los datos en forma remota es una necesidad en crecimiento. En tal sentido, un punto a tener en cuenta es el costo de los recursos de transporte necesarios para generar la comunicación. Una respuesta a esta necesidad es la propuesta de una arquitectura referente en capas ICSAD (Interfaz, Control y Sistemas de Archivos Distribuidos). La misma permite construir una implementación que facilita la descarga de los documentos y el control de versionado para el caso en el que varios usuarios estén accediendo en modo modificación.

Palabras claves: Sistemas de Archivos Distribuidos - Acceso Remoto - Sistemas Distribuidos

1. Introducción

La utilización de recursos desde distintos espacios geográficos se ha incrementado y expandido a partir del auge de los dispositivos móviles, las redes de banda ancha y las conexiones inalámbricas. Las nuevas tecnologías motivan a las organizaciones a solicitar nuevos requerimientos para mejorar la calidad y eficiencia del trabajo de sus empleados.

El acceso a recursos remotos, en especial a fuentes de datos, ha motivado una variedad de estudios [8], [4], [6], [7]. Weissman y otros [8] proponen un paradigma denominado *Smart File Object* (SFO) para alcanzar un rendimiento óptimo en el acceso a archivos remotos. En el mismo, utilizan el concepto de archivo como un tipo de objeto para proveer una interfaz de alto nivel y hacer uso de los conceptos de objetos para la invocación de las operaciones y propiedades de los archivos. El sistema de archivos *Trellis* [7], por su parte, provee una abstracción para acceder a archivos de datos utilizando nombres de archivos basados en *URL* y *SCL* y sus funcionalidades básicas están implementadas en el espacio de usuario. Una de las características que presenta es el acceso transparente a cualquier dato remoto, una capacidad importante para las áreas de metacomputación y *grid computing*.

Por otra lado, la administración de datos compartidos dentro de una empresa u organización, comúnmente se realiza a través de un sistema de archivos distribuidos o de red, con capacidades para el manejo de usuarios, permisos en una red de área local o nube propia. En algunos casos, estos sistemas no ofrecen seguridad, escalabilidad y operabilidad traspasando los límites organizacionales.

La alternativa de utilización de una red privada virtual (*VPN*) [3], en algunos casos, no es una vía de solución aceptable para las empresas, ya que no respeta las políticas de seguridad informática establecidas internamente. Cabe destacar, en este punto, que no es recomendable brindar a todos los usuarios acceso *VPN* a la red local de una organización por motivos varios de seguridad. Además, para la implementación de este tipo de accesos deben considerarse factores de rendimiento como la velocidad del enlace disponible y utilización de CPU (tanto en el cliente como en el servidor de *VPN*) según el esquema de cifrado del protocolo adoptado (PPTP, L2TP/IPSec, OpenVPN, etc.).

Miltchev y otros [6] establecen un *framework* para comparar diferentes sistemas de archivos distribuidos. En el *framework* identifican las características necesarias para esta comparación: autenticación, autorización, granularidad, delegación autónoma y revocación. Estos criterios de comparación permiten alcanzar un entendimiento de las soluciones intermedias para el acceso a datos compartidos. La relevancia de los tópicos seleccionados para la comparación está en que el acceso remoto requiere de manejo de credenciales, autorización, permisos para habilitar o no el acceso a los datos, en este caso, el acceso a los archivos.

A continuación se detallarán brevemente algunas propuestas para acceder a archivos compartidos existentes y de uso común en algunas organizaciones. Luego, más adelante, propondremos una arquitectura que pretende cubrir mayor cantidad de aspectos y funcionalidades necesarias hoy en día para el acceso a archivos compartidos desde fuera de la red interna de una organización. Para esto último presentaremos un ejemplo concreto de aplicación y detallaremos los componentes de la implementación. Por último, al final del presente trabajo se expondrán ventajas y desventajas de la propuesta presentada junto con las conclusiones del caso y desarrollos futuros.

2. Otros antecedentes

Entre las alternativas existentes para gestionar el acceso a archivos compartidos y de uso habitual en las organizaciones actuales pueden mencionarse *HFS*, mod_dir de Apache y WebDAV.

En primer lugar, *HFS* (HTTP File Server) [5] permite compartir archivos fácilmente entre un grupo de trabajo a través del protocolo *HTTP*. La compartición de los archivos se puede limitar a un grupo de usuarios o permitir que todos puedan acceder a los mismos. La diferencia con respecto a otros sistemas de archivos es que no se requiere de una red. *HFS* es un servidor web, esta característica habilita a que se publiquen los archivos a través de una presentación de un website. La utilización del protocolo *HTTP* presenta debilidades en el aspecto de seguridad, ya que el tráfico es transmitido en texto plano y cada bit de dato transportado entre el servidor web y el cliente puede ser interceptado y leído por todos los equipos que están en la cadena que pasa los datos al destino final. Esta herramienta funciona sobre el sistema operativo Microsoft Windows. Una desventaja que presenta es que no provee un esquema de autenticación a través de la interface *ADSI* (Active Directory Services Interfaces).

Una segunda alternativa, *Apache Module* mod_dir [1] se utiliza para redireccionar barra final (*trailing slash*) y servir como índice de directorio de archivos. Es una forma

simple que se utiliza para compartir archivos, en especial es usada en las fuentes de datos de libre distribución para acceder a los servidores. Una de las ventajas de esta herramienta es que se puede instalar sobre diferentes sistemas operativos sobre los cuales se puede ejecutar Apache HTTP Server como: Microsoft Windows, GNU/Linux, Unix, OS X, etc.

En tercer lugar *WebDAV* (Web-based Distributed Authoring and Versioning) es un conjunto de extensiones de HTTP, que permite a los usuarios colaborar entre ellos para editar y manejar archivos en servidores web a través de la red. *WebDAV* está documentado en RFC 2518 y extendido en RFC 3253, RFC 2518 especifica el conjunto de métodos, encabezados y tipos de contenido secundario a HTTP/1.1 para el manejo de propiedades, creación y administración de colecciones de recursos. Para usuarios comunes, *WebDAV* permite a equipos de desarrollo web y otros grupos de trabajo utilizar un servidor web remoto tan fácilmente como un servidor de archivos local.

Un ejemplo de utilización de *WebDAV* es presentado por Hernández y Pegah [2]. Para que el acceso compartido a *WebDAV* sea continuo (sin interrupciones) se le incorpora *LDAP* (Lightweight Directory Access Protocol) en el sistema para mantener una única registración. Todo esto se integra a través del servidor web *Apache* que permite la utilización de las extensiones *WebDAV* y el modelo de meta directorio *LDAP* para la autenticación de usuarios. La ventaja de esta implementación es que ofrece una solución compatible con *NFS* y OS X.

3. Arquitectura referente - ICSAD

El desafío al que intenta dar una respuesta la propuesta que presentamos en este trabajo es brindar acceso remoto desde distintos tipo de dispositivos, garantizando que se respeten las políticas de seguridad utilizadas dentro de los límites de la organización. Así la arquitectura referente, que denominamos Interfaz, Control y Sistema de Archivos Distribuidos (ICSAD), se muestra en la figura 1.

Como puede apreciarse, el modelo propuesto está organizado por capas. Cada capa es independiente, y se comunican a través de las interfaces definidas, de tal manera que si se modifica el comportamiento de las funciones no sea necesario modificar el resto de los componentes.

Los componentes principales son el sistema de archivos distribuidos (SAD), el módulo de control y el módulo de interfaz.

- Sistema de Archivos Distribuidos (SAD): este componente se encuentra dentro de los límites de la organización. Incluye todas las operaciones para el manejo de los archivos, las capacidades de acceso, la compartición de los directorios y archivos, la administración de los usuarios y permisos.
- Módulo de control: este componente es el encargado de conectar al módulo de interfaz con el sistema de archivos distribuidos, es la puerta de entrada a la organización desde el exterior. Incluye funciones para garantizar la seguridad en el acceso, permitiendo a los usuarios acceder a la información permitida desde el componente SAD. Además, se incluyen las funciones para leer, copiar, modificar, agregar documentos en el SAD.
- Módulo de interfaz: este componente se ejecuta en cada uno de los puntos de acceso remoto, como puede ser un teléfono celular, *tablet*, *notebook*, etc. Todas las operaciones requeridas sobre el SAD se realizan a través del módulo de control.

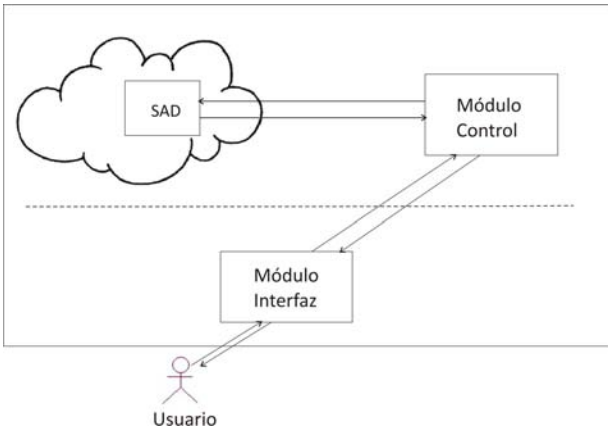


Figura 1. Arquitectura referente del Modelo ICSAD

El componente principal de esta arquitectura es el módulo de control. Que incluye las siguientes funciones:

- Lectura: la política utilizada para la implementación de esta operación es la descarga del archivo (*downloading*).
- Escritura: esta operación es equivalente a la creación de un nuevo documento en el sistema de archivos, considerando que no puede tener el mismo nombre que un archivo existente.
- Modificación: en este caso se trata de modificar un archivo existente, para ello se puede adoptar una de las siguientes políticas en el módulo de control:
  - Semántica de sesión: cuando un archivo es modificado por varios usuarios en el mismo instante de tiempo se guarda en el sistema de archivos la última copia cargada (*uploading*).
  - Semántica de versionado: en este caso, se almacenan todas las versiones del archivo bajo el mismo nombre pero con algún atributo distintivo, como puede ser el usuario, la fecha y hora de carga del archivo, o bien un identificador interno.

4. Ejemplo de aplicación

Para la arquitectura propuesta se diseña una implementación de un esquema de acceso de remoto a archivos en un repositorio distribuido ubicado sobre la red interna de una organización.

Los requisitos para la implementación que se consideraron fueron los siguientes:

- Tener acceso remoto a los archivos comunes de la organización.
- Respetar los mismos permisos de acceso sobre carpetas y archivos que brinda el SAD.
- La principal funcionalidad es el acceso en modo lectura a los archivos.
- No se podrán borrar carpetas ni archivos.

- Como funcionalidad secundaria es el acceso de escritura sobre archivos.

En el primer caso, en los repositorios se ubican los archivos compartidos, que son accesibles actualmente vía protocolo SMB (*Server Message Block*) sobre los servicios de Microsoft AD (Active Directory) de Microsoft Windows 2003R2.

Los usuarios de diferentes áreas de trabajo pueden leer aquellas carpetas y archivos sobre los que tienen permisos explícitos de acceso, otorgados según políticas de organización de la empresa. La funcionalidad principal es brindarle a los usuarios un servicio de acceso remoto a los documento internos, utilizando las mismas medidas de seguridad y políticas de acceso como si estuvieran en sus estaciones de trabajo en la red local.

Las políticas seleccionadas en el módulo de control son las siguientes: 1) para la lectura, la descarga del archivo al dispositivo; 2) para la escritura, no se permite crear un documento con el mismo nombre de otro documento en la carpeta correspondiente; 3) para la modificación, se optó por la semántica del versionado. En la selección entre las alternativas posibles para la función de modificación se consideró importante la posibilidad de brindar información a los usuarios de todas las modificaciones realizadas en un período de tiempo concurrente.

4.1. Componentes

Para la implementación de un entorno de servicio que cumpla con los requisitos funcionales especificados en la sección anterior se utilizaron los siguientes componentes:

- Máquina virtual con sistema operativo GNU/Linux distribución CentOS 6.
- Componente smbclient del producto Samba.
- Un servidor web Nginx.
- Utilización PHP-FPM para interactuar con el servidor web.
- Servicio de scripting a través del módulo PHP-CLI junto con el núcleo de PHP en su versión 5.4
- Módulo $ngx_http_ssl_module$ para proveer implementar soporte HTTPS.
- Librería OpenSSL.
- Módulo de control implementado en lenguajge PHP.
- Interface web implementada en lenguaje PHP.
- Dispositivos móviles con navegadores web.

La selección de los componentes estuvo guiada por las ventajas que ofrecen para el propósito del entorno formulado. Así, PHP es un lenguaje de *scripting*, hoy en día catalogado como de propósito general, de ejecución en servidores web. En la implementación los módulos desarrollados en PHP se ejecutan a través una interface FPM que permite un mayor rendimiento y mejores velocidades en relación a las que se hubieran alcanzado eligiendo un despliegue que utilizara el servidor web Apache y PHP cargado como un módulo del mismo.

Por su parte, PHP-FPM (FastCGI Process Manager) es una interface *FastCGI* siendo una implementación alternativa con características adicionales que hacen apropiado su uso en aplicaciones web de cualquier tamaño pero con alta cantidad solicitudes por unidad de tiempo.

En tal sentido, *FastCGI* es un protocolo que hace de interface con programas que se comunican con un servidor web. Es una variación del método de comunicación denominado *CGI* (Common Gateway Interface). El objetivo principal de este protocolo

es disminuir los tiempos adicionales asociados a la comunicación con un servidor web determinado, permitiendo al servidor web atender mayor cantidad de requerimientos al mismo tiempo.

Para proveer comunicaciones seguras entre el módulo de interfaz y el módulo de control, se optó por implementar un esquema de comunicación basado en el protocolo *HTTPS*, de esta manera se alcanzan los objetivos referentes a las políticas de confidencialidad de los datos de la organización. Además se utilizan directivas *HSTS* (HTTP Strict Transport Security) para el acceso seguro al módulo de interfaz.

La figura 2 muestra un diagrama lógico con los componentes enunciados para facilitar la comprensión del ejemplo de aplicación propuesto. En ella, la pila VM representa al módulo de Control, Clientes al módulo de Interfaz y AD al SAD conteniendo el repositorio de datos compartidos.

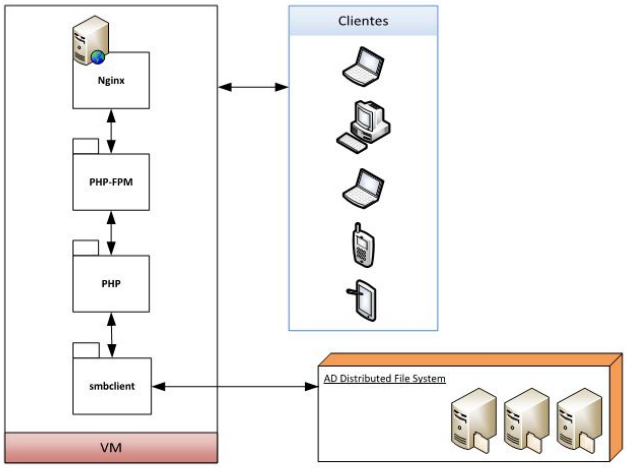


Figura 2. Diagrama de componentes de una implementación posible.

5. Conclusiones

La inserción de los medios de comunicación en la sociedad modifica la forma de efectuar las actividades cotidianas por parte de los usuarios, por ejemplo, el uso de Internet para realizar pagos, consultas, reservas, servicios de gobierno, etc. En el caso de las empresas, la conectividad a la red permite que los empleados puedan llevar a cabo sus actividades laborales desde distintas locaciones. Una alternativa es la utilización de *VPN*, que presenta ventajas y desventajas.

Como alternativa, en este trabajo se propone una arquitectura referente ICSAD para modelar el acceso a fuentes de datos que se encuentran dentro de los límites

de la organización. Esta alternativa garantiza que cada usuario obtenga las mismas capacidades y permisos como si estuviera trabajando dentro de los límites físicos de la organización.

En la arquitectura propuesta podemos definir la autenticación sobre diferentes implementaciones de archivos distribuidos que hagan uso de *LDAP* o bien algún otro tipo de sistema para catalogar usuarios y objetos con diferentes permisos, ya que el módulo de control es el encargado de la autenticación del usuario remoto.

La autorización se deja al sistema subyacente que provee dicho servicio. Así, tanto si es necesario hacer una validación sobre cada requerimiento —como en el ejemplo de implementación presentado— como si se obtiene una validación perdurable por un tiempo determinado —por ejemplo, en el caso del empleo de la API ASDI— el módulo de control puede adaptarse para ambos esquemas.

Otra ventaja de este esquema de validación de permisos es que, si el administrador modifica permisos para los objetos del sistema de archivos distribuidos, estos inmediatamente van producir un efecto en los archivos que el usuario puede o no acceder.

Por otra parte, la arquitectura provee acceso a través de la interface gráfica que se le presenta al usuario sobre cualquier dispositivo móvil mediante un navegador con soporte de protocolo *HTTPS* y lenguaje *HTML*. El objetivo es brindar un acceso universal a todos los equipos clientes sin necesidad de instalar una aplicación compilada específicamente para cada arquitectura y sistema operativo disponible hoy en día en dispositivos de uso masivo. Los usuarios no se crean sobre esta arquitectura: ya existen sobre el esquema de archivos compartidos que tenga la organización.

Además, de esta manera, se aseguran requisitos de seguridad básicos: autenticación, privacidad y auditoria. Con respecto a los dos primeros se deducen del detalle presentado en párrafos anteriores, pero podemos comentar que el tercer requisito, la auditoria, es una ventaja de la arquitectura, ya que los administradores de la organización podrán contar con un registro de todos los sucesos sobre el acceso a los archivos internos de la empresa desde el exterior. Se podrán contabilizar las autenticaciones, las descargas y modificaciones hechas por cada usuario. Cabe destacar que sin un módulo de control que provea estos servicios, en los trabajos relacionados presentados anteriormente no se dispone de toda la información necesaria para una auditoria completa, salvo que el administrador del sistema modifique compartimientos propios de dichas propuestas para incluir este tipo de auditorias, bien modificando políticas propias del sistema de archivos distribuidos (por ejemplo, en el caso de implementaciones sobre Microsoft AD) o bien *logs* de mod_dir en Apache, por citar algunos casos particulares.

Otra característica ventajosa es el hecho de que no se expone la red interna complementamente a los equipos remotos que utilizan los usuarios para acceder, como ocurre en el caso del empleo de redes privadas virtuales (VPN). Esta propuesta permite que en organizaciones que no implementan VPNs como un servicio para todos los usuarios, por sus políticas internas, se brinde este servicio en forma más general sin los inconvenientes de seguridad que presentan las VPNs. Es decir, sólo aquellos usuarios con excepciones y privilegios especiales podrán hacer uso del servicio de VPN (si estuvieran implementadas en la organización).

Como proyeccciones futuras se plantea la incorporación al módulo de control de submódulos para acceso a diferentes sistemas de archivos distribuidos, junto con la característica de autenticación sobre diversos sistemas que tengan LDAP como un servicio asociado, o bien otros esquemas de validación posibles. Por consiguiente, se aspira a poder especificar una estructura general para el módulo de control como objetivo final, para que éste, a su vez, pueda incorporar otros submódulos en la medida que resulta

necesario, a fin de ampliar el espectro de uso de la arquitectura de referencia planteada en este trabajo.

Referencias

1. The Apache Software Foundation. *Apache Module mod_dir*, 2013. httpd.apache.org/docs/current/mod/mod_dir.html.
2. L. Hernández and M. Pegah. Webdav: What it is, what it does, why you need it. In *SIGUCCS '03, ACM*, pages 249–254, 2003.
3. R. Hills. *Common VPN Security Flaws*, 2005. http://www.nta-monitor.com/.
4. T-J Liu, C-Y Chung, and C-L Lee. A high performance and low cost distributed file system. In *Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on*, pages 47–50, 2011.
5. M. Melina. *HFS Http File Server*, 2002. http://www.rejetto.com/hfs/.
6. S. Miltchev, J. Smith, V. Prevelakis, A. Keromytis, and S. Ioannidis. Decentralized access control in distributed file systems. *ACM Comput. Surv.*, 40(3):10:1–10:30, August 2008.
7. J. Siegel and P. Lu. User-level remote data access in overlay metacomputers. In *Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER'02)*, pages 1–4, 2002.
8. J. B. Weissman, M. Marina, and M. Gingras. Optimizing remote file access for parallel and distributed network applications.

XI WORKSHOP COMPUTACIÓN GRÁFICA, IMÁGENES Y VISUALIZACIÓN

- WCGIV -

CACIC2013

XI WORKSHOP COMPUTACIÓN GRÁFICA, IMÁGENES Y VISUALIZACIÓN - WCGIV -

ID	Trabajo	Autores
5631	**Chinpad, un trackpad para usuarios con discapacidades físicas**	Eliana Liberman (UNS), Emiliano Gimenez Cangelosi (UNS), Martín Larrea (UNS), Cristina Manresa Yee (UIB), Ramon Mas Sanso (UIB)
5609	**Interacción Humano Computadora para personas con capacidad motriz disminuida mediante un dispositivo Wiimote**	Brian Emanuel Megario (UNS), Matias Nicolas Selzer (UNS), Martín Larrea (UNS), Cristina Manresa Yee (UIB), Ramon Mas Sanso (UIB)
5881	**Mejorando la Conciencia Situacional en Operaciones Militares utilizando la Realidad Aumentada**	Alejandro Mitaritonna (CITEFA), María José Abásolo (UNLP)
5620	**Augmented Reality in Mobile Devices Applied to Public Transportation**	Manuel F Soto (UNS), Martín Larrea (UNS), Silvia Castro (UNS)
5807	**Aplicación turística para dispositivos móviles basada en técnicas de visión computacional**	Pablo A. Sosa (UNL), Enrique M. Albornoz (UNL), César E. Martínez (UNL)
5663	**A case study on 3D Virtual kitchen design with Kinect sensor**	Matias N Leone (UTN-FRBA), Mariano M Banquiero (UTN-FRBA), Andres Bursztyn (UTN-FRBA)
5646	**Vertex Discard Occlusion Culling**	Leonardo R Barbagallo (UTN-FRBA), Matias N Leone (UTN-FRBA), Rodrigo N Garcia (UTN-FRBA)
5718	**A Multiple Object Tracking System Applied to Insect Behavior**	Diego Marcovecchio (UNS), Natalia Stefanazzi (UNS), Claudio Delrieux (UNS), Ana Maguitman (UNS), Adriana Ferrero (UNS)

XI WORKSHOP COMPUTACIÓN GRÁFICA, IMÁGENES Y VISUALIZACIÓN - WCGIV -

ID	Trabajo	Autores
5613	**Segmentación de Imágenes de Ultrasonido por medio de un algoritmo rápido de contornos activos**	Juliana Gambini (ITBA), Ignacio Luis Bisso (UNGS)
5630	**Segmentación espectral de imágenes utilizando cámaras de tiempo de vuelo**	Luciano Lorenti (UNLP), Javier Giacomantone (UNLP)
5874	**Procesamiento de Imágenes Muestrales de Fibra Textil de Origen Animal**	Marcelo Arcidiàcono (UTN-FRC), Leticia Constable (UTN-FRC), Juan Carlos Vazquez (UTN-FRC)
5643	**Nuevos descriptores para la identificación de personas basados en la simetría del trazo**	Jorge Doorn (UNCPBA), Gladys Kaplan (UNLaM), Verónica I. Aubin (UNLaM)

Chinpad, un trackpad para usuarios con discapacidades físicas

Eliana Liberman[1], Emiliano Gimenez Cangelosi[1], Martín L. Larrea[1], Cristina Manresa-Yee[2], Ramon Mas-Sansó[2]

[1]Laboratorio de Investigación y Desarrollo en Visualización y Computación Grafica (VyGLab), Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, ARGENTINA
[2]Universitat de les Illes Balears. Unidad de Gráficos, Visión por Computador e Inteligencia Artificial, ESPAÑA
el@cs.uns.edu.ar, emilianogimenezcangelosi@hotmail.com, mll@cs.uns.edu.ar, {cristina.manresa, ramon.mas}@uib.es

Abstract. Las computadoras son instrumentos fundamentales para personas con discapacidades físicas ya que posibilitan compensar funciones disminuidas o ausentes. Las mismas facilitan la participación de personas con discapacidades en todos los niveles de la vida social, cultural y económica. Constituyen un medio por el cual personas con discapacidades pueden comunicarse, estudiar y recrearse equiparando oportunidades. En el presente trabajo se ataca la problemática de desarrollar un medio que permita a una persona incapacitada de mover cualquier parte del cuerpo con excepción de la cabeza, utilizar una computadora. Se busca así, brindar a usuarios que están cautivos en un cuerpo sin movilidad que no les permite expresarse y/o comunicarse adecuadamente, una herramienta para que puedan hacerlo a través de una computadora.
Keywords: Discapacidad, Computadora, Interacción humano-computadora, Interfaces táctiles.

1 Introducción

La informática brinda muchas facilidades en diversos niveles a las personas con discapacidades físicas. Por un lado, constituyen un medio de comunicación mediante el cual una persona con incapacidad de habla y escritura puede expresarse con otras personas. Por otra parte, equiparan las oportunidades que alguien discapacitado puede alcanzar, dado que permiten que un individuo que está incapacitado de hablar y escribir pueda participar de clases educativas y posteriormente de una carrera universitaria.
Como se dijo, la informática permite la participación de las personas con discapacidad en todos los niveles de la vida social, cultural y económica ([1,2,3,4,5]). Por esta razón es de actual preocupación facilitar el uso de las computadoras a personas discapacitadas existiendo diversas opciones en el mercado como se detallará más adelante. Para atacar esta problemática, se analizó, diseñó y desarrolló un trackpad

que pueda manejarse utilizando el mentón el cual se denominó “Chinpad” (“Chin” por “mentón”).
La estructura de este artículo es la siguiente: A continuación se brindará una reseña de las opciones existentes en el mercado actual que atacan la problemática descripta. Luego, en la Sección 3, se procederá a introducir nuestra propuesta, describiendo sus ventajas y desventajas como así también una comparación con las opciones desarrolladas previamente. La sección 4 describe los detalles de implementación de nuestra propuesta. En la sección 5 se incluye la experiencia tenida con los alumnos de la escuela especial 509 y el feedback que se obtuvo de la misma. Finalmente, este artículo concluye delineando las conclusiones y presentando el trabajo a futuro.
Para el desarrollo del presente proyecto, se trabajó en conjunto con la escuela de chicos especiales de Bahía Blanca número 509. Se tuvieron en cuenta consideraciones explicadas por las docentes profesionales del mismo y se tomaron como parámetro dos niños alumnos de la institución con discapacidades físicas que están aprendiendo a utilizar una computadora.
Los mismos son incapaces de mover cualquier parte del cuerpo con excepción de la cabeza, y hasta el momento se comunican con la pc mediante diversos dispositivos y técnicas que se detallarán en la siguiente sección.
Tomando en cuenta tanto las ventajas como las desventajas de esta interacción detalladas por las docentes a cargo, se decidió construir un dispositivo específico para permitir manejar con mayor facilidad una computadora valiéndose solo de movimientos suaves de la cabeza.
Esto es, se desarrolló el Chinpad para un usuario target con las siguientes características: Inmovilidad del cuerpo con excepción de la cabeza. Capacidad de efectuar movimientos de cabeza controlados (permitiendo un factor pequeño de movimientos involuntarios).

2 Trabajo Previo

La presente sección describe las alternativas más relevantes relacionadas con el tema de estudio de este artículo. Los trabajos aquí descriptos cubren soluciones tanto de hardware como de software. Cada alternativa será presentada junto con sus ventajas, desventajas y costo.

2.1 Mouse por barrido y conmutador auxiliar

El mouse por barrido ([6,7]), es un software que provee una interfaz que alterna periódicamente entre todas las opciones de un cursor (click, doble click, click derecho, arrastrar, mover derecha, mover izquierda, etc). Con el pulsador, se efectúa un click sobre la opción deseada para obtener el comportamiento indicado con el cursor. Se tiene como ventajas la capacidad de emulación total de las características de un mouse y una curva de aprendizaje baja. Las desventajas son una interacción usuario computadora muy lenta (debido al retardo entre las distintas opciones del

cursor) y una falta de independencia usuario-computadora (otro individuo debe configurar la PC para utilizar el software por barrido y luego el usuario con discapacidad puede iniciar su uso). Esta alternativa corresponde a las tecnologías utilizadas en la escuela 509. El costo estimativo de un pulsador corriente es de €60[1]. Existen otros tipos de conmutadores que permiten detectar más de una acción. Algunas alternativas se detallan a continuación:

El conmutador doble, para lengua, mentón o mejilla ([8,9]), que consiste en una varilla y un doble sensor que permite detectar dos sentidos de movimiento. Está diseñado para ser accionado con esfuerzos mínimos y recorridos muy cortos de la lengua, el mentón, la mejilla u otras partes del cuerpo. Las ventajas son por un lado que permite detectar el movimiento en dos sentidos, por lo que se contaría con dos posibles estados a detectar a diferencia de un pulsador en el que solo contamos con uno y además, incluye un *driver* para configurar entre otras cosas la sensibilidad de detección de los movimientos. Como desventajas, decimos que la palanca es frágil, está pensada para captar movimientos reducidos con mucha sensibilidad. Movimientos bruscos podrían dañarla. Por otro lado, la varilla podría introducirse accidentalmente en la boca. El precio de este conmutador es de €95[1].

Otra alternativa es el conmutador de doble soplido y aspiración ([10]), esté cuenta con un tubo y dos salidas de conmutador. Al soplar por el tubo se activa una de las salidas de conmutador y al aspirar por el mismo tubo se acciona la otra salida permitiendo el control de dos funciones diferentes. Se debe colocar el tubo en la boca y cerrar bien los labios a su alrededor. Entre sus ventajas se destacan que permite detectar dos acciones: soplido y aspiración. Además, incluye un *driver* para configurar entre otras cosas la fuerza necesaria para la activación. Las desventajas a mencionar son que puede producirse la obstrucción del tubo por saliva por lo que debe limpiarse habitualmente, también puede producirse la obstrucción del tubo por pliegues. Además, conlleva un esfuerzo que puede cansar al usuario. Su precio es de €145[1]. Una tercera alternativa constituye el conmutador con sensor de parpadeo de fibra óptica ([11,13]), este dispositivo posee una correa que permite colocar una fibra óptica cerca de un ojo, y es capaz de detectar el parpadeo. Las ventajas son que se adapta a cualquier tipo de discapacidad, ya que no requiere de un esfuerzo motriz grande. A su vez, el sensor ignora los parpadeos involuntarios, solo considera aquellos que duren tres segundos o más. Entre sus desventajas podemos encontrar que el *delay* de los parpadeos no es configurable, puede producirse la obstrucción del cable por pliegues y conlleva un esfuerzo que puede cansar al usuario. Por otra parte, resulta más agobiante que utilizar un pulsador común y dificulta la visión. Su precio es de usd 645[1]. La última alternativa que analizaremos se trata del conmutador con sensor de movimientos musculares ([12]), dicho conmutador es un dispositivo que posee dos sensores circulares los cuales se activan con pequeños movimientos musculares, como por ejemplo levantar una ceja o apretar la mandíbula. También se lo puede activar con el cambio de temperatura que provoca la respiración o el tacto. Sus ventajas son que se adapta a cualquier tipo de discapacidad, ya que no requiere de un esfuerzo motriz grande. Además, permite detectar dos acciones diferentes y un

[1] Precio tomado en el año 2013.

control por hardware permite ajustar la sensibilidad para discriminar acciones involuntarias, así como un tiempo de retardo. Su precio es de usd325[1].

2.2 Sistemas basados en visión

Una alternativa son los dispositivos que a través de una o más cámaras capturan el movimiento de ciertas partes del cuerpo y lo interpretan. Estos sistemas incluyen un software específico para emular las distintas funcionalidades de un mouse dependiendo del movimiento capturado. Analizaremos dos alternativas existentes en el mercado IRISCOM y EnableViacam. IRISCOM ([14]), es un dispositivo que captura el movimiento de los ojos. Con el movimiento del iris se controla el cursor y con el pestañeo se hace la acción de click. Sus ventajas son que se adapta a cualquier tipo de discapacidad, ya que no requiere de un esfuerzo motriz. No se requiere control cefálico. Posee una interfaz en la cual se elige la acción del próximo pestañeo, click derecho, arrastrar y soltar, doble click o click izquierdo. Entre sus desventajas podemos encontrar que la experiencia ha demostrado que existen dos tipos de molestias que pueden ocasionarse por usar aparatos de seguimiento del ojo, estas son, daños en el cuello (el aparato requiere restringir el movimiento de la cabeza, lo cual con el tiempo puede causar molestias en el cuello) y molestias en los ojos (mientras se use el sistema los músculos de sus ojos se ejercitan de una forma muy precisa y controlada). Otra desventaja presente es que el tiempo que una persona puede usar el sistema antes de sentir molestias o fatiga varía de un usuario a otro. Por lo tanto, cada uno debe sentir cuando es el momento para descansar. Por último, cuenta con un precio elevado y el software se cobra por separado. Su precio es de €6000 hardware y €1800 software + instalación[1]. EnableViacam ([15]), es un software que permite controlar el puntero del mouse simplemente moviendo la cabeza. El primer paso consiste en calibrar el programa, para que identifique los movimientos de la cabeza, luego de lo cual el puntero responderá a estos movimientos. Para realizar los clicks, sólo es necesario dejar el puntero relativamente quieto durante un período de tiempo configurable. Entre sus ventajas encontramos que, permite emular completamente el comportamiento del mouse. Además, la velocidad del puntero, la aceleración y suavizado, el tiempo de detención, y otras variables pueden ser configuradas para ajustarse a las necesidades del usuario. A su vez, está diseñado específicamente para eliminar la necesidad de asistencia una vez instalado (autonomía). Su desventaja es que requiere de un esfuerzo motriz muy grande para mover toda la cabeza. Otra gran ventaja es que es gratuito.

2.3 Sistemas basados en el movimiento del mentón

Por último abordaremos algunas alternativas de dispositivos que al igual que nuestra propuesta emulan el comportamiento del mouse utilizando el mentón. Evaluaremos dos alternativas existentes en el mercado la primera de ellas diseñada especialmente para personas discapacitadas y la segunda no. La primera alternativa se trata de BJoy Mentón. Este dispositivo es un mouse tipo “joystick”. El reducido tamaño de la palanca, la disposición de sus botones y sus opciones de sujeción facilitan su

utilización con el mentón. Se conecta a la computadora a través del puerto USB. Sus ventajas son que provee de todas las funcionalidades de un mouse y permite autonomía. Además, incluye un *driver* para configurar la función de cada botón, la velocidad, la orientación y otros parámetros. A su vez, los ajustes se almacenan en el dispositivo que posee una memoria interna. Su desventaja es que esta indicado solo para usuarios que tienen un buen control de movimiento de la cabeza. Demanda un esfuerzo motriz grande ya que para mover el cursor se necesita mover una palanca analógica. Su precio es de €500[1]. La segunda alternativa que analizaremos será Logitech Wireless Touchpad. Esta alternativa es desarrollada por Logitech y utiliza un touchpad externo inalámbrico. Las ventajas de esta alternativa son que provee de todas las funcionalidades de un mouse. Además, permite autonomía y posee dos botones para los clicks. Sus desventajas son que no está implementado para reconocer el tamaño de una barbilla sino para detectar dedos, por lo que al querer utilizarlo con el mentón, a veces algún movimiento no será detectado y/o deberá realizarse algún esfuerzo importante para utilizarlo apoyando una pequeña porción de la misma. Por otro lado, los botones se encuentran delante del pad táctil por lo que resulta incómodo, y hasta en algunos casos, imposible que el usuario los presione. Su precio es de usd 35[1].

3 Propuesta

A partir de un análisis realizado sobre las distintas alternativas mostradas en la sección dos, se encaró una nueva propuesta.
Se buscó un hardware que pudiera ser utilizado fácilmente por sus usuarios finales, sin dificultades de aprendizaje y que permitiera alcanzar la misma velocidad que un mouse común.
Se buscó además, alcanzar independencia total en el uso del dispositivo generando un software que se inicie al encender la pc y que no requiera de configuración previo a su uso.
De esta manera, la propuesta de este proyecto fue analizar, diseñar y desarrollar un trackpad que pueda manejarse utilizando el mentón con el objetivo de facilitar la interacción de personas con cierta discapacidad motriz con la PC, permitiéndoles manejar a voluntad el cursor con el movimiento de su mentón. Se decidió que sea manejado con el mentón dado que resultó la parte más conveniente de la cabeza. Con la nariz o la frente no se podría mirar hacia la computadora mientras se utiliza el Chinpad. La aplicación fue basada en la idea de las cajas táctiles. En la figura 1(a) se muestra el dispositivo final.
En líneas generales, mediante movimientos de la barbilla sobre la superficie de trackpad, una cámara web situada dentro capta dichos movimientos, los cuales son luego traducidos en el movimiento del cursor en la pantalla.
Para manejar ambos clicks (derecho e izquierdo) se incluyó un mouse a uno de los lados del trackpad de manera tal que los botones estén a la altura de la superficie del mismo y puedan ser presionados con facilidad sin necesidad de levantar mucho la

barbilla. El mismo cuenta con botones redondos y separados que son fáciles de apretar con un movimiento hacia al costado del mentón.
Por sugerencia de personas involucradas con chicos que presentan estas discapacidades, se le dio a la superficie del trackpad el mismo tamaño, aproximadamente, a un trackpad presente en una notebook. Esto se debe a que un tamaño superior involucraría un mayor movimiento de cuello lo cual generaría un mayor esfuerzo en estos chicos.
La caja presenta el mínimo tamaño que fue posible realizar teniendo en cuenta el tamaño de la cámara utilizada que debe caber dentro de la caja. A su vez, fue necesario considerar que la cámara necesita estar a una cierta distancia mínima para visualizar la superficie completa. Es claro que el costo de esta alternativa consiste mayormente en el costo de una cámara web sencilla, lo cual es muy económico.
En resumen, la propuesta presentada cuenta con las siguientes características: Es un dispositivo que permite emular íntegramente el comportamiento de un mouse a través del movimiento del mentón del usuario sobre una superficie y de botones para poder hacer clicks. Consiste de una caja de madera cubierta de una tapa semitransparente, que contiene una cámara en su interior para captar las sombras producidas por el movimiento de la barbilla del usuario y de dos botones en uno de sus costados que pueden ser presionados fácilmente con la propia barbilla. Sus ventajas son que permite manejar el cursor con mayor fluidez que la mayoría de las alternativas existentes, que tiene una curva de aprendizaje muy leve, no requiere de una configuración previa para su uso permitiendo autonomía a los usuarios y resulta mucho más económica que el resto de las alternativas. Además, la caja puede inclinarse y su grado de inclinación puede ser regulado según las necesidades de cada usuario.
Sus desventajas son que debe utilizarse en una habitación iluminada con una luz de normal a suave (no muy fuerte), y el hecho de que sea bastante aparatoso (debido al tamaño considerable del dispositivo y su peso). El soporte debe insertarse en algún lugar donde quede firme y tanto la inclinación de la caja como la altura del brazo deben ajustarse al usuario antes de poder utilizar el dispositivo. Su costo es de aproximadamente 60 dólares[1].

4 Implementación

4.1 Software

Para capturar las imágenes tomadas por la cámara y filtrarlas hasta detectar sólo la sombra del mentón, apoyada en una determinada posición sobre el vidrio, se utilizó el software Community Core Vision(CCV) ([16,17])(este se configuró con los siguientes valores: Se seleccionaron los campos *show outlines*, *show ids*, *use camera*, *inverse*, *TUIO udp*, *fingers*, *smooth*, *highpass*, *Amplify*. Se asignaron valores específicos en los campos *image threshold* (249), *movement filtering* (0), *min blob size* (8), *max blob size* (793), *Smooth* (6), *Blur* (110), *Noise* (27), *Aplify* (32)). En el CCV se activaron las opciones de tracking y de envió de paquetes TUIO con la

información de los objetos trackeados. Estos paquetes son capturados por un programa desarrollado, usando el lenguaje JAVA. Este software utiliza la librería "libTUIO" para conectarse, mediante un puerto determinado, al servidor de paquetes TUIO (en este caso el CCV) e interpretarlos. Hay dos formas de mover un cursor con una superficie táctil. Una, siguiendo un estilo que denominaremos normal y otra siguiendo lo que denominaremos estilo trackpad. Pueden encontrarse en otra bibliografía los términos absoluto y relativo. La opción elegida fue la de mover el cursor al estilo de un trackpad. Siguiendo el modo normal, cada porción del trackpad se corresponde con la porción correspondiente a la misma posición en la pantalla. Esto es, el cursor se mueve a la posición mapeada en pantalla según la posición donde se colocó la barbilla sobre la superficie. En cambio, en el estilo trackpad, el movimiento del cursor continua desde la última posición que se registró al remover la barbilla de la superficie, sin importar dónde la misma es colocada nuevamente. La clase principal, ejecutable, implementa la interface TUIOlistener provista por la librería, por lo que posee métodos para manejar los eventos de aparición de un nuevo objeto, movimiento del mismo y eliminación. Para implementar el movimiento del cursor se utilizó la clase Robot predefinida de java y se implementó en cada uno de estos métodos la acción del cursor correspondiente. A su vez fue un requerimiento de las docentes de los chicos de las escuelas especiales, que los mismos pudieran adquirir la mayor autonomía posible para utilizar la PC. Esto es, no requerir de una configuración y activación detalladas que sean necesarias realizar antes de permitirles a los usuarios disponer del trackpad para controlar la pc. Por esta razón, se creó un ejecutable que corre al inicio de la máquina de forma que con solo encenderla el trackpad es activado y la PC está lista para ser utilizada por los usuarios.

4.2 Hardware

Se utilizó una cámara web estándar. En particular se usó la cámara Microsoft LifeCam VX-1000 dado que se contaba con ella. No es necesario que se trate de una cámara infrarroja a diferencia de otras de las alternativas mencionadas en la sección 2. Esto es un aspecto interesante por dos razones. Por un lado, las cámaras infrarrojas son más difíciles de conseguir y por el otro, una cámara web estándar claramente es más económica. Se adquirió un mouse con botones grandes, redondos y separados para realizar los clicks como se dijo anteriormente. Se construyó la caja con madera y la superficie se realizó con un vidrio. Entre el vidrio y la cámara hay un papel en blanco para permitir que la misma solo capte la sombra del mentón. Se decidió colocar el papel del lado de adentro (y no del lado de afuera) del vidrio para permitir que la sombra se propague de forma más atenuada a la cámara, dado que con la cabeza y el cuello se genera una sombra mayor a la generada al momento de realizar una caja táctil para utilizar con los dedos. Para su uso es necesario permanecer en una habitación iluminada pero con una luz de normal a suave. Es preferible que no sea muy fuerte. Con respecto a esto, se incluyó en el dispositivo una pequeña lámpara que acompaña a la luz de la habitación. La misma será utilizada o no dependiendo de la iluminación del ambiente, es decir, representa una ayuda que puede o no ser necesaria. Por otro lado, el trackpad fue pensado para ser utilizado de forma inclinada.

Esto es, que la superficie del trackpad no esté horizontal. Resultó más cómodo construirla inclinada hacia el usuario dado la discapacidad motriz del mismo. De todas maneras, el grado de inclinación puede ser regulado como se mostrará más adelante en algunas imágenes. Por último, la posición de la cámara puede ser ajustada hasta alcanzar la calibración adecuada.

4.3 Sobre el armado

Antes de construir el trackpad final, fueron probados distintos tamaños para la caja. En un comienzo la misma era más grande (se utilizó una caja de zapatos), brindando una superficie mayor y una distancia grande entre la cámara y la misma. Dado que fue construida en cartón, se utilizaron elementos domésticos (libros, vasos, papel de cocina, etc) para darle soporte ya que la porción de vidrio aún sin cortar era muy pesada. En la figura 3 se puede apreciar dos fotos de ese prototipo. A partir de este prototipo, se realizaron varias pruebas de modelos intermedios hasta llegar al tamaño óptimo, que fue el menor tamaño logrado.

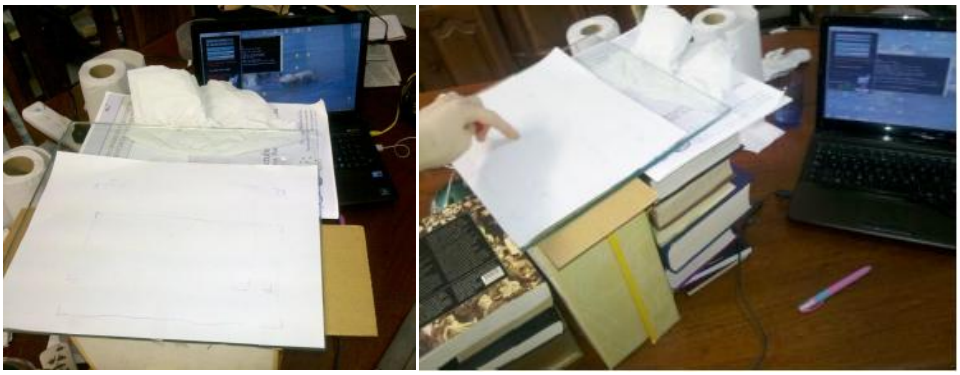


Figura 3. Prototipos iníciales Chinpad

(a) Dispositivo final previo a la experiencia con usuarios.

(b) Dispositivo final posterior a la experiencia con usuarios.

Figura 1. Últimas dos versiones dispositivo final.

5 Experiencia con usuarios

El dispositivo fue probado por tres alumnos de la escuela especial 509 con distintos problemas motrices. El primero de ellos, debido a que poseía movimientos involuntarios que no le permitían desplazar con facilidad el mentón sobre la superficie, no pudo realizar las pruebas con el dispositivo. Los restantes pudieron utilizarlo y se familiarizaron con él en poco tiempo, con cierta ayuda en un principio. Durante nuestra experiencia, los inconvenientes que se nos presentaron fueron sobre todo durante la instalación del dispositivo. Hubo que buscar una habitación iluminada y buscar un soporte donde apoyarlo, que fuera resistente, que se pudiese ajustar a una altura donde los alumnos pudieran llegar a él con su mentón y que no dificultara el ingreso de su silla de ruedas por debajo, al acercarse.
Por esto último, se decidió añadir al dispositivo un soporte que sea lo suficientemente resistente y regulable a la vez y que permita acercar el Chinpad a los usuarios de manera cómoda y sin estorbar el paso de la silla de ruedas. En la figura 1, se puede visualizar las dos versiones del dispositivo, una anterior a la experiencia con los usuarios y una posterior.

6 Conclusiones y Trabajo a futuro

El acceso a una computadora, y por su intermedio a Internet, brinda un amplio mundo de posibilidades, tanto para la recreación como para la educación y comunicación. En este artículo se presentó una solución que tiene como beneficios ser de bajo costo y de simple construcción. Los próximos pasos a seguir en este trabajo serán continuar capacitando a los alumnos de la escuela 509 para que aprendan a utilizar el dispositivo con mayor fluidez, difundir la existencia de esta alternativa liberando el código fuente de la implementación bajo la licencia GPLv3 y finalmente donar una o más unidades del dispositivo a la escuela en cuestión.

7 Agradecimientos

Este trabajo fue parcialmente financiado por el proyecto español MAEC-AECID FRIVIG A1/037910/11 y el proyecto argentino 24/N028 de la Secretaría General de Ciencia y Tecnología de la Universidad Nacional del Sur

8 Referencias

1. A.F. Newell, P. Gregor (2002, November). Design for older and disabled people – where do we go from here?. Department of Applied Computing, The University of Dundee, Nethergate, Dundee, DD1 4HN, United Kingdom, GB.
2. Batya Friedman (Ed.). (1997). Human values and the design of computer technology (No. 72). Cambridge University Press.

3. Mussa-Ivaldi, F. A., Casadio, M., & Ranganathan, R. (2013). The body-machine interface: a pathway for rehabilitation and assistance in people with movement disorders. Expert Review of Medical Devices, 10(2), 145-147.
4. Murata, Y., Yoshida, K., Suzuki, K., &Takahashi, D. (2013, February). Proposal of an Automobile Driving Interface Using Gesture Operation for Disabled People. In ACHI (2013), The Sixth International Conference on Advances in Computer-Human Interactions (pp. 472-478).
5. Abascal, J., Garay, N., & Gardeazabal, L. Sistemas de Interacción Persona-Computador para usuarios con discapacidad. INTERACCIÓN 2.000.
6. Zappalá, D., Köppel, A., & Suchodolski, M. (2011). Inclusión de TIC en escuelas para alumnos con discapacidad motriz.
7. Pavón Rabusco, f., & Ordóñez Sierra, r. (1999). Las nuevas tecnologías como recursos de apoyo para el aprendizaje de las personas con necesidades educativas especiales. Cabero j, y otros coordinadores. Nuevas tecnologías en la formación flexible y a distancia. Sevilla: Edutec.
8. Salem, C., & Zhai, S. (1997, March). An isometric tongue pointing device. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 538-539). ACM.
9. Bates, R. (2002, March). A computer input device selection methodology for users with high-level spinal cord injuries. In Proceedings of the 1st Cambridge Workshop on Universal Access and Assistive Technology (CWUAAT); 25th-27th March.
10. Iga, S., & Higuchi, F. (2002). Kirifuki: Inhaling and exhaling interaction with visual objects. Entertainment Computing: Technologies and Applications, (pp. 133-140).
11. Jacob, R. J. (1993). Eye movement-based human-computer interaction techniques: Toward non-command interfaces. Advances in human-computer interaction, 4, 151-190.
12. Rantanen, V., Niemenlehto, P. H., Verho, J., & Lekkala, J. (2010). Capacitive facial movement detection for human–computer interaction to click by frowning and lifting eyebrows. Medical & biological engineering & computing, 48(1), 39-47.
13. Lin, M., & Mo, G. (2011, October). Eye gestures recognition technology in Human-computer Interaction. In Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on (Vol. 3, pp. 1316-1318). IEEE.
14. Clemotte, A., Raya, R., Ceres, R., & Rocon, E. (2013). Preliminary Result from a Multimodal Interface for Cerebral Palsy Users Based on Eye Tracking and Inertial Technology. In Converging Clinical and Engineering Research on Neuro rehabilitation (pp. 443-448). Springer Berlin Heidelberg.
15. Nowosielski, A., & Chodyła, Ł. (2013, January). Touchless Input Interface for Disabled. In Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013 (pp. 701-709). Springer International Publishing.
16. http://ccv.nuigroup.com/ Web oficial del Community Core Vision
17. http://sethsandler.com/multitouch/community-core-vision-guide/ Guía de configuración de Community Core Vision

Interacción Humano Computadora para personas con capacidad motriz disminuida mediante un dispositivo Wiimote

Brian Emmanuel Magario[1], Matias Nicolás Selzer[1], Martín L. Larrea[1], Cristina Manresa-Yee[2], Ramón Mas-Sansó[2].

[1]Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica (VyGLab), Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Argentina.
{matias.selzer, brian.magario}@gmail.com
mll@cs.uns.edu.ar

[2]Universitat de les Illes Balears. Unidad de Gráficos, Visión por Computador e Inteligencia Artificial, España.
Ed. Anselm Turmeda, Crta Valldemossa km 7.5, 07122 Palma, España.
{cristina.manresa, ramon.mas}@uib.es

ABSTRACT. En nuestra era moderna, las computadoras son la base fundamental para poder estar conectados con el resto del mundo. Este derecho no se le debe privar a nadie, ni siquiera a aquellas personas con capacidades motrices disminuidas. En los últimos años, interfaces de todo tipo han sido desarrolladas en el área de la Interacción Humano Computadora para ayudarles a acceder a una computadora de la forma más segura y eficiente posible.
En este artículo se presenta una interface cómoda y sobre todo económica para que personas con capacidad motriz disminuida en sus brazos y manos puedan controlar satisfactoriamente el mouse del ordenador, abriéndoles paso a un mundo infinito de posibilidades.

Palabras Clave: Head Tracking, Interacción Humano Computadora, Wiimote, Discapacidad Motriz.

1. Introducción

Las personas que poseen algún tipo de discapacidad motriz a menudo tienen grandes dificultades para comunicarse o interactuar con los dispositivos electrónicos y tecnológicos de la actualidad. Existen herramientas de asistencia tecnológica que se han desarrollado para ayudarlos a superar este tipo de situaciones, brindándoles la posibilidad de comunicarse mediante programas y hardware especializados. De esta forma, este tipo de usuarios pueden beneficiarse del acceso a una computadora, ya sea para obtener conocimientos, actividades recreacionales y uso de internet.

Para las personas con capacidades motrices disminuidas o nulas en brazos o manos, el uso de los movimientos de la cabeza, de ser posibles, abre un abanico de posibilidades.

El principal medio de interacción con una computadora es el ratón y el teclado. Hoy en día coexisten muchas alternativas para el control del mismo. Por ejemplo, es posible encontrarse con algunos sistemas que usen tecnología infrarroja; u otros, en donde el usuario pueda mover el cursor mediante movimientos de su cuerpo. Una dificultad o desventaja de estos dispositivos es que, en muchos casos, existen usuarios que poseen impedimentos propios de la discapacidad que traen aparejados movimientos involuntarios, provocando la necesidad de una recalibración, siendo necesaria la presencia de otro individuo para poder realizar dicha actividad. Para controlar esto se utilizan sitios donde apoyar la pera, que pueden ser incómodos. Desafortunadamente los dispositivos existentes a la fecha suelen ser extremadamente caros o con un funcionamiento ineficiente o inexacto.

Nuestro objetivo ha sido desarrollar un dispositivo confiable, económico y cómodo de usar para que personas con incapacidad motriz en los brazos o manos puedan hacer uso de una computadora.

En este artículo se presentará una descripción de las diferentes alternativas existentes hasta la fecha, seguido de nuestra propuesta, y posteriormente su implementación. Concluiremos el artículo realizando comparaciones correspondientes y una conclusión al respecto.

2. Trabajos Relacionados

Existen diversos dispositivos y sistemas en el mercado para tratar la comunicación entre personas con alguna discapacidad y las computadoras.

Dispositivos de adquisición y procesamiento de señales neurológicas ([1,2]), como el Emotiv EEG Neuroheadset (Figura 1.a) pueden utilizarse para el control del mouse del ordenador. Estos dispositivos censan las señales eléctricas producidas por el cerebro del usuario para detectar sus pensamientos y expresiones. Como estos dispositivos no fueron creados para simplemente controlar un mouse, la desventaja es que mucho de su poder computacional se desperdicia. Son dispositivos altamente sofisticados y costosos (alrededor de u$s750.00). Además, es mucho el tiempo que requiere el usuario para adaptarse al mismo y aprender a utilizarlo correctamente.

Otro tipo de sistemas utilizan la cámara web de la computadora para procesar la imagen del usuario y que éste pueda así controlar el mouse de la computadora mediante movimientos de su cabeza ([3,4]). La ventaja de esto es que no se requiere de hardware adicional, más que la cámara web integrada en cualquier notebook de hoy en día. Además, existen sistemas como Camera Mouse que son totalmente gratis.

El problema en estos sistemas es la poca precisión de los mismos, el usuario debe tener mucha precaución a la hora de mover su cabeza si desea un correcto movimiento del mouse en la pantalla, por lo que no está pensado para todo el público. Además necesita condiciones externas muy estrictas para funcionar adecuadamente, como por ejemplo una luz adecuada y ningún obstáculo entre el usuario y la cámara en ningún momento, ya que esto último desconfigura la calibración del sistema completamente. El SINA ([5]) es un sistema que utiliza una cámara web y mediante avanzadas técnicas de software realiza el seguimiento del usuario.

Existen también sistemas con cámaras especiales con la capacidad de detectar

pequeños puntos infrarrojos que el usuario debe colocarse en la frente o en los anteojos. Estos sistemas, como Madentec Tracker 2000 o Tracker Pro poseen un buen desempeño a la hora de controlar el mouse ya que fueron diseñados para eso. Al utilizar hardware adicional, el alto costo de estos dispositivos (u$s 950.00) se convierten en su mayor desventaja. Además, el hecho de tener que pegarse unos pequeños puntos en la frente no suele ser de mucha comodidad para el usuario.

Los Sistemas de Visión por Computadora Centralizados en el Usuario utilizan la última tecnología en cuanto a cámaras y métodos de procesamiento de imágenes y calibración muy sofisticados. Estos sistemas suelen ser más caros ya que poseen un hardware especial, pero la ventaja es su funcionalidad, ya que pueden detectar perfectamente la pupila del usuario para que el mismo pueda mover el mouse en su computadora con tan solo mover el ojo ([6,7]). Estos dispositivos están pensados para personas con un muy bajo grado de movilidad, pero puede ser utilizado por todo el público en general.

Existen otros sistemas, como Jouse 2 (Figura 1.b), el cual utiliza un joystick para que el usuario pueda controlar el mouse de su computadora mediante su boca. Aunque el precio sea extremadamente alto (u$s 1,499.99) y la comodidad para el usuario no sea la mejor, es uno de los dispositivos más precisos a la hora de controlar el mouse. Otros dispositivos como NoHands MouseTM consiste en dos pedales separados operados con los pies, con uno se controla el puntero y con el otro los clicks. El precio de este dispositivo no es muy alto (u$s 360.00), pero la adaptación que debe hacer el usuario para acostumbrarse a utilizarlo suele ser tediosa y poco intuitiva. Además, los resultados no son tan buenos ni precisos, como otros dispositivos existentes en el mercado.

(a) Emotiv EEG Neuroheadset

(b) Jouse 2

Figura 1. Ejemplos de productos existentes.

Tras haber analizado todas las propuestas existentes, nuestro enfoque se basa en la creación de un dispositivo económico y óptimo, con completa funcionalidad y comodidad para todo tipo de usuarios, pero enfocándonos especialmente en aquellas personas con alguna discapacidad física que les impida utilizar un ratón común para el uso cotidiano del ordenador.

3. Nuestra Propuesta

El objetivo general de nuestro trabajo fue desarrollar un dispositivo que sirva como interfaz entre el usuario y la computadora permitiéndole de esta manera controlar la misma de una forma análoga a la utilización física del mouse. Principalmente el dispositivo debe ser cómodo de utilizar, eficiente, intuitivo, robusto, y económico.

Nuestra idea se enfoca en que el usuario, para poder mover el puntero del ratón, deba rotar levemente la cabeza hacia la dirección deseada. Se han elegido movimientos de rotación por ser más suaves y más intuitivos y requieren menos esfuerzo que los movimientos de desplazamiento.

El dispositivo Wiimote está provisto de una cámara infrarroja con la que puede detectar la posición de puntos infrarrojos. Utilizando este dispositivo y unas gafas especiales provistas de tres LEDs infrarrojos colocados estratégicamente, se detectan los movimientos de la cabeza del usuario para poder mover el cursor en la pantalla del ordenador. Mediante una conexión Bluetooth se envía luego la posición relativa de dichos LEDs a nuestra aplicación corriendo en el ordenador. En la figura 2 podemos apreciar la posición de dichos LEDs infrarrojos sobre las gafas.

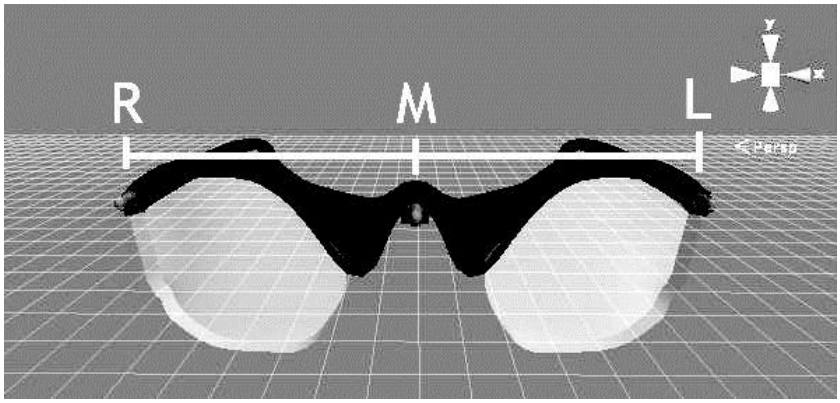


Figura 2. Imagen frontal del dispositivo

Llámese L al LED izquierdo, R al derecho y M al del centro. Podemos obtener la siguiente expresión matemática referida a las distancias entre los LEDs:

$$|R - M| \cong |L - M| \quad (1)$$

Aprovechando la facilidad del Wiimote para capturar la emisión de los LEDs infrarrojos, se procesa el movimiento del puntero según las distancias relativas entre los LEDs de la siguiente forma:

Cuando el usuario gire la cabeza hacia la izquierda o hacia la derecha, la distancia entre L e R con respecto a M variarán. Si el usuario gira su cabeza hacia la izquierda,

$$|R - M| > |L - M| \quad (2)$$

Análogamente, cuando el usuario gire su cabeza hacia la derecha,

$$|R - M| < |L - M| \tag{3}$$

De esta forma podremos conocer cuándo mover el puntero hacia la izquierda o hacia la derecha. En la figura 3 se puede apreciar claramente la diferencia de distancias entre los distintos LEDs cuando el usuario gira su cabeza hacia la izquierda.

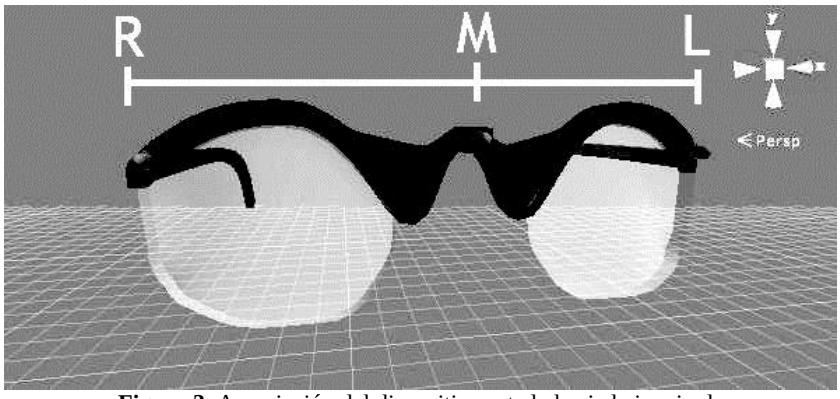


Figura 3. Apreciación del dispositivo rotado hacia la izquierda

Si siguiéramos el mismo criterio, serían necesarios dos LEDs más situados arriba y abajo del LED central para poder diferenciar los movimientos del usuario hacia arriba y hacia abajo respectivamente. Esto puede ser una solución válida, pero poco práctica teniendo en cuenta el poco espacio físico del dispositivo. Para evitar este problema, se plantea la siguiente solución: el LED del medio se ubicará unos centímetros más adelante que los LEDs de los extremos, como se puede ver en la figura 4.

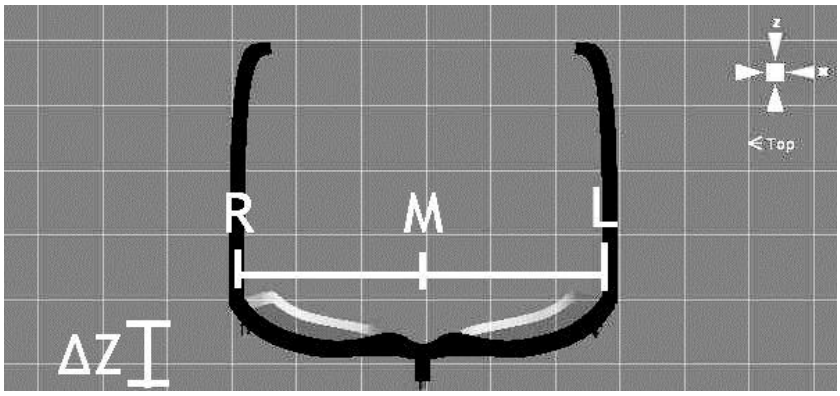


Figura 4. Vista superior del dispositivo

Para medir los movimientos hacia arriba o hacia abajo, se calcula un promedio entre la posición de los LEDs de los extremos sobre el eje Y, y éste valor se compara

con el valor sobre el eje Y del LED central. El promedio del valor de los LEDs de los extremos se realiza porque se asume que, cuando el usuario inclina su cabeza hacia arriba o hacia abajo, la posición de éstos será aproximadamente la misma, y no estaremos discriminando entre uno u otro.

Entonces, si

$$\frac{(R_y + L_y)}{2} < M_y \tag{4}$$

Esto es, si el LED central está más arriba que el promedio entre los LEDs de los extremos, el usuario está inclinando su cabeza hacia arriba y por lo tanto el puntero se moverá hacia esa dirección. Podemos ver esto en la figura 5.

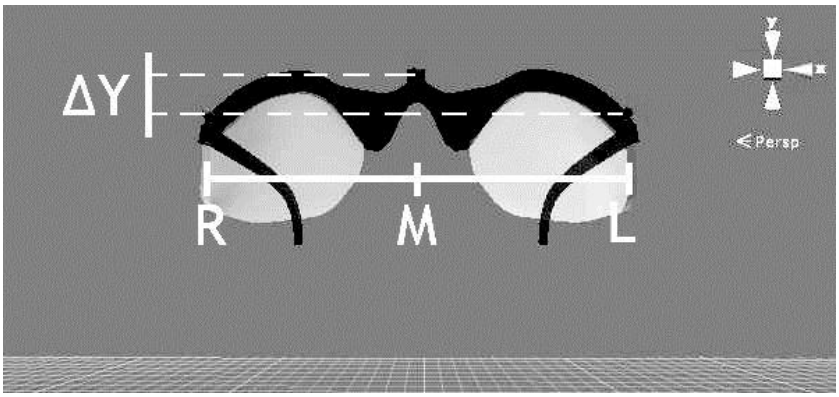


Figura 5. Movimiento hacia arriba del dispositivo

$$\Delta Y = \left| \frac{(R_y + L_y)}{2} - M_y \right| \tag{5}$$

Caso contrario, si

$$\frac{(R_y + L_y)}{2} > M_y \tag{6}$$

El usuario está inclinando su cabeza hacia abajo, moviendo el puntero respectivamente. Aprovechando la facilidad de que el LED central se encuentra más adelante que los otros, este generará un ángulo de giro más grande con un movimiento más leve por parte del usuario, es decir, se acercará más rápidamente a los LEDs de los extremos y por ende, las distancias entre los mismos serán más notorias, facilitando su obtención y la realización de cálculos.

En nuestra implementación, la velocidad con la que se moverá el puntero por la pantalla no será siempre la misma, sino que será relativa a las distancias entre los LEDs. Es decir, cuanto más incline la cabeza el usuario, más rápido se moverá el puntero. Esta aceleración será verdaderamente intuitiva para el usuario, ya que deberá realizar un mayor movimiento para generar una mayor velocidad.

Por último, para que el puntero no esté siempre moviéndose por la pantalla y el usuario pueda dejarlo quieto, se plantea un umbral a la hora de comparar las

distancias entre los LEDs. Por ejemplo, si la distancia entre L y M es casi la misma que la distancia entre R y M, o la diferencia de esas distancias es menor al umbral, significa que el usuario tiene su cabeza situada prácticamente de frente, por lo que el puntero no se moverá. Analíticamente, el puntero se mantendrá quieto siempre que se cumpla:

$$|(|R - M| - |L - M|)| < umbral \quad (7)$$

Con respecto a los eventos del mouse, se quiso llegar a una solución en donde no se deba utilizar absolutamente nunca los brazos o las manos, por lo que se optó por utilizar comandos de voz. El usuario podrá realizar los eventos típicos del mouse, como lo son el click, click derecho, doble click, arrastrar y soltar, mediante comandos simples de voz.

4. Implementación

Se ha desarrollado una aplicación en C# para el análisis y procesamiento de los datos recolectados por dispositivo Wiimote, a partir de los cuales se realiza el cálculo correspondiente de la posición y velocidad del mouse en la pantalla. En la figura 6 puede verse una vista de la aplicación en cuestión, donde se aprecian los LEDs de las gafas encendidos; en la figura 7 puede verse a un usuario utilizando la aplicación.

Para facilitar la interpretación de dichos datos, el código provisto por Johnny Chung Lee ([8]) en sus investigaciones relacionadas a la interacción humano computadora con el dispositivo Wiimote, ha sido de gran utilidad.

Con respecto a los eventos del mouse realizados mediante comandos de voz, se hizo uso de la librería System.Speech provista por el sistema operativo Windows. Esta librería brinda la posibilidad de almacenar los comandos deseados en un archivo de texto para luego ser interpretados durante la ejecución de la aplicación.

Tanto el código de la aplicación como las instrucciones de armado de las gafas se encontrarán disponibles en el siguiente sitio web bajo licencia freeware:
https://docs.google.com/file/d/0Bynuc25YJ_y-Zm5hcm55UU5GLTQ/edit?usp=sharing

Figura 6. Se aprecian los LEDs del dispositivo encendidos.

Figura 7. Usuario utilizando la aplicación.

5. Conclusión

Se ha desarrollado un dispositivo que, en conjunto con el existente Nintendo Wiimote y una aplicación de pc, permite a personas con cierto grado de discapacidad motriz en sus brazos y manos poder controlar el mouse, brindándole acceso a la tecnología mediante el uso de las computadoras. Las experiencias han sido muy alentadoras, demostrando que nuestro sistema puede proveer un adecuado y cómodo acceso a una computadora a personas con cierto grado de discapacidad motriz, de una forma muy satisfactoria.

6. Trabajo a Futuro

En un futuro se trabajará a nivel de implementación y algoritmo para generar un mejor movimiento del puntero del ratón, más suave y adecuado para el usuario.

Por otro lado, en cuanto a las conexiones entre el Wiimote y la computadora, se buscará que sean manejadas automáticamente por la aplicación en cuestión, evitando así la dependencia de aplicaciones externas.

Agradecimientos. Este trabajo fue parcialmente financiado por el proyecto español MAEC-AECID FRIVIG A1/037910/11 y el proyecto argentino 24/N028 de la Secretaría General de Ciencia y Tecnología de la Universidad Nacional del Sur.

Referencias

1. Lusted, H. S., & Knapp, R. B. (1996). Controlling computers with neural signals. Scientific American, 275(4), 82-87.
2. Moore, M. M. (2003). Real-world applications for brain-computer interface technology. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 11(2), 162-165.
3. Varona, J., Manresa-Yee, C., & Perales, F. J. (2008). Hands-free vision-based interface for computer accessibility. Journal of Network and Computer Applications, 31(4), 357-374.
4. Akram, W., Tiberii, L., & Betke, M. (2007). A customizable camera-based human computer interaction system allowing people with disabilities autonomous hands-free navigation of multiple computing tasks. In Universal Access in Ambient Intelligence Environments (pp. 28-42). Springer Berlin Heidelberg.
5. Cristina Suemay Manresa Yee, Junio 2009, Advanced and natural interaction system for motion-impaired users, Tesis Doctoral, Universitat de les Illes Balears, Departament de Ciències Matemàtiques i Informàtica,España.
6. Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. Mind, 2(3), 4.
7. Kaufman, A. E., Bandopadhay, A., & Shaviv, B. D. (1993, October). An eye tracking computer user interface. In Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in (pp. 120-121). IEEE.
8. Johnny Chung Lee, HCII, johnny@cs.cmu.edu, Website: http://johnnylee.net/projects/wii/.

Mejorando la Conciencia Situacional en Operaciones Militares utilizando la Realidad Aumentada

Alejandro Mitaritonna[1], María José Abásolo[2,3]

[1] Instituto de Investigaciones Científicas y Técnicas para la Defensa (CITEDEF)
San Juan Bautista de La Salle 4397 (B1603ALO) Villa Martelli, Buenos Aires, Argentina
amitaritonna@citedef.gob.ar

[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA)
[3] Instituto de Investigación en Informática LIDI (III-LIDI)
Facultad de Informática – Universidad Nacional de La Plata (UNLP)
calle 50 y 120 (1900) La Plata, Buenos Aires, Argentina
mjabasolo@lidi.info.unlp.edu.ar

Resumen. Durante las operaciones militares, los campos de batalla se convierten en zonas fracturadas donde el nivel de confusión, el ruido y la ambigüedad impactan en la manera de alcanzar los objetivos tácticos. La Conciencia Situacional (CS) se convierte en un reto ya que la percepción de la situación es inestable, lo que conduce a la comprensión degradada y a la incapacidad del soldado en proyectar los resultados apropiados. Para afrontar dicho reto diversos proyectos militares han centrado sus esfuerzos en diseñar un sistema digital integrado como soporte para la toma de decisiones del personal militar en ambientes desconocidos. En particular, este trabajo presenta una recopilación actualizada de algunos sistemas digitales que utilizan la Realidad Aumentada (RA) como un medio para la representación visual de la información adquirida del contexto. Adicionalmente, se propone un framework de RA cuyo objetivo es mejorar la CS de los soldados en el campo de batalla mediante el uso de la RA.

Palabras claves: Realidad Aumentada, Conciencia Situacional, Dispositivos Móviles, Conocimiento del Contexto, Guerra Centrada en Redes

1 Introducción

De acuerdo a Bryant, D. et al [3] el fratricidio sigue siendo una amenaza muy real en los campos de batalla actuales. Para hacer frente a esto, los militares han puesto mucho esfuerzo en el desarrollo de tecnologías de identificación en combate para mejorar la capacidad de los soldados en identificar al enemigo con precisión. Saarelainen, Tapio et al [19] afirman que las futuras operaciones militares se basarán

en herramientas de Comando, Control, Comunicaciones, Informática, Información, Inteligencia (en inglés C4I2, Command, Control, Communications, Computers, Information, Intelligence) para un rendimiento óptimo en sus tareas asignadas en ambientes versátiles y hostiles.

La CS es una representación mental y comprensión de los objetos, eventos, interacciones, condiciones ambientales y cualquier otro tipo de factores de una situación específica que puedan afectar al desarrollo de las tareas humanas. Muchas de las operaciones militares se desarrollan en entornos desconocidos. Las soluciones de CS permiten a los soldados hacer un uso efectivo de la información variada en un contexto de batalla siendo uno de los principales objetivos la reducción de la carga cognitiva en momentos de stress. Las nuevas tecnologías ofrecen métodos innovadores de obtener información contextual y representarla visualmente de una manera natural y no invasiva sin afectar el proceso cognitivo del combatiente. Es el caso de la Realidad Aumentada (RA).

La RA, definida por Azuma, R. [1] se refiere a aplicaciones interactivas en tiempo real donde se visualiza la realidad con elementos sintéticos agregados (objetos 3D, sonidos, texto, etc.). Existen diversos proyectos que incorporan el uso de RA en aplicaciones militares, ya que con su uso podría producir mejoras dramáticas en el rendimiento del soldado y proporcionar una gran ventaja en el combate.

El resto del artículo se organiza de la siguiente forma: la sección 2 introduce definiciones como Conciencia Situacional, Conocimiento del Contexto, Cognición Aumentada, Realidad Aumentada y Guerra Centrada en Redes. La sección 3 presenta una revisión de diferentes proyectos militares que utilizan la RA para mejorar la CS en el campo de batalla. La sección 4 se propone un framework de software de RA. Por último la sección 5 presenta las conclusiones y trabajos futuros.

2 Definiciones

2.1 Conciencia Situacional

Brown, David Wm [2] menciona que la CS (en inglés, situation awareness o también situational awareness) se refiere a la percepción, la comprensión, y la previsión de los elementos dentro de un entorno operacional requerido para actuar con eficacia dentro de ese ambiente.

Tremblay, Sébastien et al [22] definen que la CS es un requisito previo para la oportuna y correcta toma de decisiones en el rápido y altamente estresante contexto de los entornos operativos de infantería. Se espera que la introducción de las tecnologías de soporte electrónico en el campo de batalla mejore la CS, proporcionando la información correcta, en el momento adecuado y en el formato correcto.

Por otra parte Endsley, M. R [7], [8] menciona que la CS es la percepción de los elementos en el medio ambiente dentro de un volumen de tiempo y espacio, la comprensión de su significado y la proyección de su situación en un futuro próximo.

En la figura 1 se grafica el modelo de la CS en la toma de decisión dinámica.

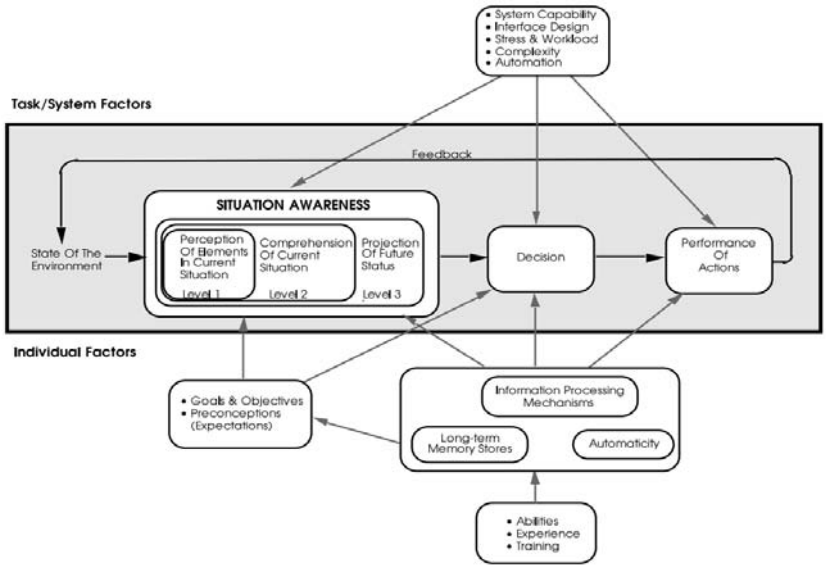


Fig. 1 – Modelo de la CS en la toma de decisión dinámica (Endsley, 1955)

Endsley, M. R. et al [9] determinan que uno de los factores más importantes que subyacen en el desarrollo de una adecuada CS es la presencia de los modelos mentales y esquemas de situaciones prototípicas. Proporcionan una construcción mental fundamental para dirigir la forma de asignar la atención y destacar los temas críticos.

2.2 Conocimiento del Contexto

La definición formal del Conocimiento del Contexto (en inglés, context-aware) más ampliamente aceptada ha sido proporcionada por Dey, A. y Abowd, G. D. [6]: "*El contexto es cualquier información que se puede utilizar para caracterizar la situación de una entidad. Una entidad puede ser una persona, un lugar o un objeto que se considera relevante para la interacción entre un usuario y una aplicación, incluido el usuario y las propias aplicaciones*".

Para Dey, Anind K. et al [5] el contexto puede ser considerado como un conjunto de información que incluye la actividad del usuario, ubicación, preferencias personales y el estado actual. La movilidad crea situaciones en las que el contexto del usuario, tales como la ubicación de un usuario y las personas y objetos a su alrededor, es dinámico y se va modificando. El contexto se define mejor como estados o configuraciones ambientales, tales como: ubicación, orientación, tiempo, objetos cercanos o personas, nivel de luz ambiental, ruido y temperatura.

De acuerdo a Schilit, B. et al [21] al proveer acceso al contexto se incrementa la riqueza de la comunicación hombre-máquina y la efectividad de la elaboración de la tarea.

Hull, R. et al [12] las define como los sistemas de cómputo capaces de sensar, interpretar y responder de acuerdo con el entorno en que se encuentra el usuario.

2.3 Cognición Aumentada

Para desarrollar un sistema de visualización de información se debe examinar cuáles son las necesidades de información y determinar cuál es la mejor modalidad o combinación de modalidades que consistiría en presentar esa información con el fin de que el sistema sea robusto, utilizable y eficaz. La capacidad de procesamiento de información de los seres humanos se ha convertido rápidamente en un factor limitante en la interacción hombre-máquina. Este problema ha motivado el desarrollo de una nueva disciplina científica mencionada en [15] llamada Cognición Aumentada (CA). Métodos para detectar y mitigar las limitaciones de procesamiento humano de la información y el diseño de soluciones para mejorar el intercambio y uso de la información en los sistemas hombre-máquina son las preocupaciones específicas de la CA.

2.4 La Realidad Aumentada

De acuerdo a Hicks, Jeffrey et al [10] la RA proporciona al usuario información superpuesta que se puede visualizar en el mundo real, es decir, complementa al mundo real con información virtual. La RA mejora la percepción del mundo natural mediante el agregado de información a los sentidos, ya sean visuales, sonidos, olores o sensaciones táctiles. La RA se refiere a la mezcla de las señales virtuales a partir del entorno tridimensional real en la percepción del usuario. Denota la fusión 3D de imágenes sintéticas en la visión natural del usuario del mundo circundante, utilizando gafas o un HMD (en inglés, head-mounted display). A través de la capacidad de presentar la información superpuesta, integrados en el entorno del usuario, la RA tiene el potencial de proporcionar beneficios significativos en muchas áreas de aplicación. Muchos de estos beneficios surgen del hecho de que las señales virtuales presentadas por un sistema de RA pueden ir más allá de lo que es físicamente visible.

2.5 Guerra Centrada en Redes

De acuerdo al DoD [4] la guerra centrada en redes (en inglés, Network-centric warfare) es una doctrina militar que apunta a convertir una ventaja informativa en una ventaja competitiva mediante una sólida red de fuerzas, geográficamente dispersas, pero bien conectadas e informadas.

Moffat, J. y Atkinson, S. R. [17] describen que se está dirigiendo hacia una estructura organizacional de Guerra Centrada en Redes que es plana, rápida y está basada en la información, en contraste con la estructura jerárquica de movimiento lento, basado en el modelo de comando y control que ha definido los sistemas de gestión del siglo 20. En la Guerra Centrada en Redes, las computadoras integran la información adquirida a partir de múltiples fuentes, para aumentar la CS del espacio de batalla en tres dimensiones y crear una imagen que proporciona información crítica y relevante para todos los niveles de mando y control, que incluyen al soldado. Las redes se forman a través de nodos con la información transmitida a través de los puestos de mando, vehículos, y computadoras portátiles de los soldados.

3 Revisión de proyectos militares

3.1 Antecedentes

Zieniewicz, Matthew J. et al [24] mencionan que en el año 1989, el Ejército de EE.UU. utilizó una pequeña computadora portátil para ayudar a los soldados en las tareas de campo de batalla.

James Schoening, analista de investigación que trabajó en el CECOM (Communications Electronics Command) del Ejército de EE.UU., es quien comienza a utilizar computadoras portátiles. Trabajando con Matt Zieniewicz, Schoening transformó su idea en una arquitectura de sistema con tecnologías específicas, tales como la transmisión inalámbrica de datos, captura de imágenes y Sistema integrado de Posicionamiento Global (GPS). En 1990, Schoening y Zieniewicz se asociaron con John Flatt, Sal Barone y Almon Gillette para demostrar el sistema *Soldier's Computer*. Más tarde, basándose en el proyecto Soldier's Computer, dio origen al proyecto SIPE (Soldier Integrated Protective Ensemble). El proyecto SIPE, dirigido por Carol Fitzgerald, fue el primero en que el Ejército de EE.UU trató a los diversos componentes de los equipos de combate como un sistema integrado.

3.2 Proyecto Eyekon

Hicks, Jeffrey et al [10] definen al proyecto EyeKon como un sistema de soporte a la toma de decisiones basado en agentes inteligentes instalados en una computadora portátil que transporta el soldado. El soldado visualiza información de objetivos y otra información en su armamento. El proyecto tiene como objetivo desarrollar iconos inteligentes y descripciones que se superponen en el video del arma del soldado. Las funciones básicas de Eyekon se encuentran en una computadora portátil conectada vía una red inalámbrica segura a una base de datos local y remota. Incorpora sensores que brindan información en tiempo real (por ejemplo sensor inercial, GPS, IR, etc). El software está compuesto por agentes que realizan consultas a la red a fin de monitorear información de amenazas y otros tipo de información estratégica para el soldado (por ejemplo misión, estados, etc). Sobre la pantalla del arma se sobreimprime información utilizando técnicas de RA.

3.3 Proyecto BARS

El Naval Research Laboratory (NRL) desarrolló un sistema prototipo de realidad aumentada conocido como BARS (en inglés, Battlefield Augmented Reality System) [16]. Este sistema conecta a múltiples usuarios móviles junto con un centro de mando. BARS se centró en el desarrollo de un sistema digital para ayudar a resolver el creciente énfasis en las operaciones militares en terreno urbano. BARS realiza el seguimiento (tracking) de la posición y orientación de la cabeza del usuario y superpone gráficos junto con anotaciones que se alinean con los objetos reales en el campo visual del usuario. Varias unidades compartían una base de datos común, donde las personas podían optar por unirse a un canal determinado para acceder a los datos gráficos.

3.4 Proyecto iARM

En el año 2009, el Defense Advanced Research Projects Agency (DARPA) contrató a la empresa Tanagram Partners para desarrollar el proyecto Intelligent Augmented Reality Model (iARM) [14]. Básicamente, el objetivo de iARM es desarrollar un sistema digital integrado que podría mejorar significativamente la toma de decisiones del personal militar en entornos complejos a través de un sistema operativo integrado, un modelo de servicios de datos, y un HMD mejorado. El objetivo es que todos estos componentes trabajen juntos de una manera transparente permitiendo a los soldados percibir, comprender y, lo más importante, proyectar el mejor curso de acción para un mayor rendimiento para alcanzar los objetivos tácticos. El proyecto iARM abarca muchos de los atributos de la inteligencia artificial. En la figura 2 se muestra el diseño conceptual del HMD y la visión del soldado a través de las gafas.

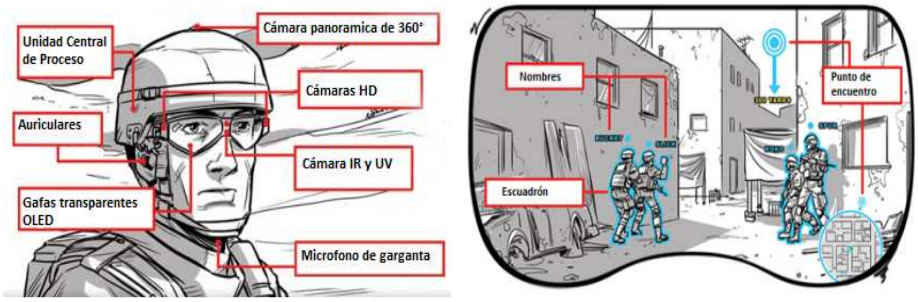


Fig. 2 – Diseño conceptual del proyecto iARM

3.5 Proyecto ULTRA-Vis

En [23] se detalla que el programa Urban Leader Tactical Response, Awareness and Visualization (ULTRA-Vis), soportado por el DARPA en su fase 1, ha desarrollado un prototipo de un sistema de RA para los soldados en el campo de batalla. El sistema ULTRA-Vis superpone iconografía gráfica a todo color en la escena local observada por el soldado. El programa desarrolló e integró un sistema de poco peso, una pantalla *see-through* holográfica de bajo consumo con un sistema de visión de tracking de posición y orientación. Usando el sistema ULTRA-Vis, un soldado puede visualizar la ubicación de otras fuerzas, los vehículos, los peligros y las aeronaves en el medio ambiente local, incluso cuando éstos no son visibles para el soldado. Además, el sistema puede ser utilizado para comunicar al soldado de una variedad de información tácticamente significativa incluyendo imágenes, rutas de navegación y alertas. El prototipo estará dotado para el reconocimiento gestual mediante el uso de un guante. Permitirá sobreimprimir símbolos en el campo de batalla en 3D, localizar objetivos enemigos y ubicar a las fuerzas aliadas. ULTRA-Vis provee a los escuadrones una ventaja táctica muy clara permitiendo la colaboración entre los integrantes del escuadrón. Posibilitará una alta conciencia situacional y la habilidad de tomar decisiones mientras se está en movimiento en el campo de operaciones. En la figura 3 se muestra el diseño conceptual del proyecto ULTRA-Vis.

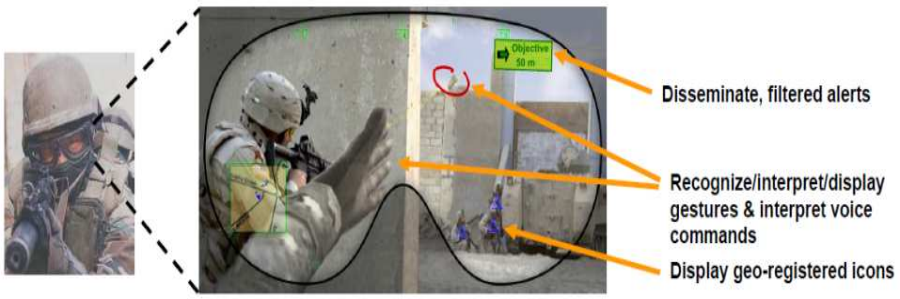


Fig. 3 – Diseño conceptual del prototipo ULTRA-Vis

4 Framework de RA propuesto

En esta sección se presenta el framework RAIOM (Realidad Aumentada para la Identificación de Objetivos Militares). La subsección 4.1 detalla el denominado Modelo de Información - desde la obtención de la información, su procesamiento, hasta su visualización. La subsección 4.2 explica la importancia del filtrado y representación de la información obtenida del contexto a través de la transformación que va teniendo la información hasta la representación visual final. Teniendo en cuenta lo antes mencionado y luego de haber analizado los proyectos militares detallados en la sección anterior, la subsección 4.3 describe el framework RAIOM, el cual utiliza a la RA como tecnología de representación visual y cuyo fin será mejorar la CS de los soldados en operaciones militares.

4.1 Modelo de Información

Al proceso de transformación de la información obtenida del contexto para representarla visualmente de una manera natural la hemos denominado *Modelo de Información* (Fig. 4). Dicho modelo define como la información pasa por diferentes etapas hasta la representación visual de la misma. Estas etapas se centran en *Adquirir*, *Enviar*, *Procesar* y *Representar* la información del contexto. La etapa de *Adquirir* denota la obtención de información del contexto, principalmente, a través de sensores dispersos geográficamente. La etapa de *Enviar* corresponde al envío de la información adquirida en la etapa anterior por medio de componentes de comunicación. La etapa de *Procesar* se encarga de computar la información obtenida del contexto para luego enviarla a un dispositivo móvil que se encargará de tratar a la información mediante técnicas de pre-procesamiento, detección, extracción, clasificación, reconocimiento, identificación, etc. La última etapa del modelo se refiere a *Representar* la información que fue procesada en la etapa anterior. La representación de la información es visual y se utilizan técnicas como la RA para enriquecer la percepción del usuario del mundo real.

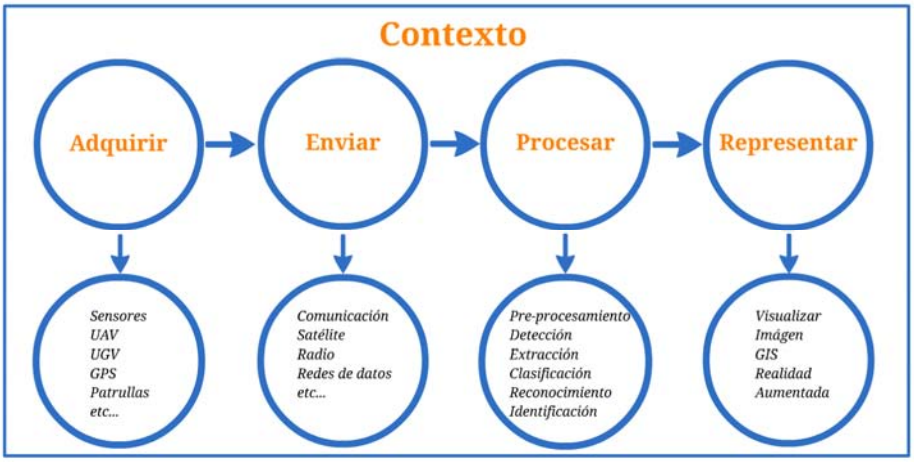


Fig. 4 – Etapas del Modelo de Información

4.2 Representación y filtrado de la información

Julier, S. et al [13] han presentado la idea de utilizar el contexto del mundo real para obtener información, implementando un sistema que filtra la información basada en la ubicación física, superponiendo información mediante el uso de la RA. El objetivo principal de filtrar información es priorizar y reducir la cantidad de información que se presenta con el fin de mostrar sólo lo que es relevante para el usuario.

Sestito, S. et al [20] mencionan que debido a la movilidad del soldado a través del medio ambiente, el contexto puede cambiar drásticamente dependiendo de su posición. La cantidad de información que se puede mostrar a un usuario en un mundo virtual puede ser abrumadora. Para paliar este problema, el sistema debe ordenar y priorizar la información de modo tal que se deben mostrar las características que son "más relevantes" para el soldado, como por ejemplo amenazas.

4.3 Proyecto RAIOM

Habiendo analizado los proyectos militares que utilizan la RA para mejorar la CS de los soldados en los campos de batalla, proponemos el diseño de un framework de software que contemple las mejores características de los mencionados proyectos y se ajusten a las necesidades concretas del personal militar. El objetivo del proyecto RAIOM, es mejorar la CS del combatiente tomando información del contexto y representarla visualmente mediante el uso de la RA para ayudar al soldado a tomar decisiones bajo presión. El proyecto RAIOM se basa en una comprensión del estado actual de la tecnología digital, la naturaleza cambiante del combate, la evolución del papel del soldado y la creciente importancia de la CS.

Capacidades Operativas:

- Interacción mediante reconocimiento gestual y vocal para la toma de datos
- Tracking múltiples (GPS, sensores, visión, etc.)
- Detección y reconocimiento de objetos tridimensionales
- Reconocimiento facial

- Identificación de aliados y enemigos
- Filtrado de información prioritaria
- Implementación e integración del prototipo en dispositivos móviles

Características:

- Autónomo (poca dependencia de acceso a la red externa)
- Omnidireccional (comunicación entre los integrantes de la patrulla y el centro de Comando y Control)
- Liviano (reconocimiento Gestual/Vocal) y componentes de bajo consumo.
- Seguro (cifrado a Canal + Datos –Proyecto C-RAIOM-)
- Código abierto (Framework + SO)
- Móvil (Smartphones, tabletas y gafas)

5 Conclusión y trabajos futuros

En este artículo se ha descripto como se puede mejorar la CS utilizando la RA como una técnica avanzada de representación de la información contextual en un entorno bélico. Para un mejor entendimiento se definió el significado de términos como Conciencia Situacional, Conocimiento del Contexto, Realidad Aumentada, Cognición Aumentada y Guerra Centrada en Redes. Se hizo una revisión histórica de los sistemas digitales militares que utilizan la RA para mejorar la CS. Se describió el Modelo de Información para explicar cómo la información va pasando por diferentes etapas hasta llegar al último eslabón del modelo que es la representación visual de la información. Se detalló la importancia del filtrado de la información ya que dicta qué se debe mostrar al soldado y cuándo hacerlo. La investigación presentada sirve para reunir y analizar el estado del arte de los proyectos militares de cara a realizar el diseño del framework de software denominado RAIOM. Dicho framework será diseñado para brindar soporte para la toma de decisiones del personal militar en ambientes desconocidos utilizando la RA.

Referencias

1. *Azuma R.* (1997). A survey of Augmented Reality. Presence: Teleoperators and Virtual Environments, vol. 6, no. 4, pp. 355-385
2. *Brown, David Wm.* (2012). A Survey of Mobile Augmented Reality Technologies for Combat Identification Applications. MSc thesis. Athabasca University.
3. *Bryant, D.; Smith, D.* (2009). Comparison of Identify-Friend-Foe and Blue-Force Tracking Decision Support for Combat Identification. DRDC: Toronto, Rep. 2009-214
4. *Department of Defense of USA –DoD-* (2005). The Implementation of Network-Centric Warfare. Washington, D.C.. p. 7
5. *Dey, Anind K.; Abowd, Gregory D.; Brown, Peter J.; Davies, Nigel; Smith, Mark; Steggles, Pete* (1999). Towards a Better Understanding of Context and Context-Awareness. Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing. Pages 304-307.
6. *Dey, A.; Abowd G. D.* (2000). Towards a better understanding of context and context-awareness. En: CHIA'00 workshop on Context-Awareness.

7. *Endsley, M. R.* (1988). Design and evaluation for situation awareness enhancement. In Proceeding of the Human Factors Society 32nd Annual Meeting (pp. 97-101). Santa Mónica, CA: Human Factors Society.
8. *Endsley, M. R.* (1995). A taxonomy of situation awareness errors. In R. Fuller, N. Johnston & N. McDonald (Eds.), Human Factors in Aviation Operations (pp. 287-292). Aldershot, England; Averbury Aviation, Ashgate Publishing Ltd.
9. *Endsley, M. R.; Bolstad, Cheryl A.; Jones, Debra G.; Riley, Jennifer M.* (2003). Situation Awareness Oriented Design: From User's Cognitive Requirements to Creating Effective Supporting Technologies. Human Factors and Ergonomics 47th Annual Meeting, Denver, Colorado, EEUU.
10. *Hicks, Jeffrey; Flanagan, Richard; Dr. Petrov, Plamen; Dr. Stoyen, Alexander* (2003). Eyekon: Distributed Augmented Reality for Soldier Teams. © Copyright 21st Century Systems, Inc.
11. *Holmquist, J.; Barnett, J.* (2001). Digitally Enhanced Situation Awareness: An Aid to Military Decision-Making. Proceedings of the Human Factors and Ergonomics Society Annual Meeting , vol. 45 no. 4 542-546
12. *Hull, R.; Neaves, P.; Bedford-Roberts, J.* (1997). Towards situated computing. En: 1st International Symposium on Wearable Computers, pp. 146–15
13. *Julier, S.; Lanzagorta, M.; Baillot, Y.; Rosenblum, L.; Feiner, S.; Hollerer, T.; Sestito S.* (2000). Information filtering for mobile augmented reality. In: Augmented Reality. (ISAR 2000). Proceedings. IEEE and ACM International Symposium.
14. *Juhnke, Joseph; Kallish, Adam; Delaney, Dan; Dziedzic, Kim; Chou, Rudy; Chapel, Tim.* (2010). Tanagram Partners. Final Project Report. Aiding Complex Decision Making through Augmented Reality: iARM, an Intelligent Augmented Reality Model.
15. *Kobus, D. A.; Brown C. M.* (2006). DARPA Improving Warfighter Information Intake Under Stress—Augmented Cognition. Pacific Science & Engineering Group, Inc. SSC San Diego.
16. *Livingston, Mark A.; Rosenblum, Lawrence J.; Julier, Simon J.; Brown, Dennis; Baillot, Yohan; Swan II, J. Edward; Gabbard, Joseph L.; Hix, Deborah* (2002). An Augmented Reality System for Military Operations in Urban Terrain. Proceedings of Interservice / Industry Training, Simulation & Education Conference (I/ITSEC), December 2 -5, Orlando, Florida, page 89 (abstract only).
17. *Moffat, J.; Atkinson, S. R.* (2002). Libro: The Agile Organization: From Informal Networks to Complex Effects & Agility.
18. *Moon, Yong-Woon; Jung, Hae-Sun; Jeong, Chang-Sung* (2010). Context-awareness in Battlefield using Ubiquitous Computing. Network Centric Warfare. 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010)
19. *Saarelainen, Tapio; Jormakka, Jorma* (2010). C4I2-Tools for the Future Battlefield Warriors. IEEE - Fifth International Conference on Digital Telecommunications.
20. *Sestito, Sabrina; Julier, Simon, Lanzagorta, Marco; Rosenblum, Larry* (2000). Intelligent Filtering for Augmented Reality. In: Proceedings of SimTecT 2000, Sydney, Australia.
21. *Schilit, B.; Adams, N.; Want R.* (1994). Context-aware computing applications. En: 1st International Workshop on Mobile Computing Systems and Applications, pp. 85-90.
22. *Tremblay, Sébastien; Jeauniaux, Patrick; Romano, Paul; Lowe, Jacques; Grenier, Richard* (2011). A Multi-Perspective Approach to the Evaluation of a Portable Situation Awareness Support System in a Simulater Infantry Operation. IEEE - International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), Miami Beach, FL.
23. *ULTRA-Vis* (2008). BAA 08-36. Broad Agency Announcement for Information Processing Techniques Office and Defense Advanced Research Projects Agency.
24. *Zieniewicz, Matthew J.; Johnson, Douglas C.; Wong, Douglas C.; Flatt, John D.* (2002). The Evolution of Army Wearable Computers. PERVASIVE Computing.

Augmented Reality in Mobile Devices Applied to Public Transportation

Manuel F. Soto[1], Martín L. Larrea[2], and Silvia M. Castro[2]

[1] Instituto de Investigaciones en Ingeniería Eléctrica (IIIE) "Alfredo Desages"
Univesidad Nacional del Sur,
Consejo Nacional de Investigaciones Científicas y Técnicas.
[2] Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica (VyGLab),
Departamento de Ciencias e Ingeniería de la computación,
Universidad Nacional del Sur.
{mfs,mll,smc}@cs.uns.edu.ar

Abstract. Augmented Reality (AR) is one of the most revolutionary technologies at these times. It improves the real world by additional computer generated information. The AR paradigm opens new ways for development and innovation of different applications, where the user perceives both, virtual and real objects at the same time. With the rise of SmartPhones and the development of its characteristics, the AR on mobile devices emerging as an attractive option in this context. In this paper we present the design, implementation and testing of an application of AR on Android platform for mobile devices. This allows a person traveling through the city gets information of routes, timeouts, etc; about a particular bus line. All this information is provided on the mobile device and associated to the real world, facilitating their interpretation.

Keywords: Augmented Reality, OpenGL ES, Android, Public Transport, OpenStreetMap

1 Introduction

Nowadays, technological advances in the area of mobile devices are constant. The increase in processing power, the storage, quality of cameras and screens have given rise to the development of applications of Augmented Reality (AR) on these devices. This situation is further benefit by the low cost of mobile devices, and easy Internet access from them. In this context, we designed and developed an application to assist the user that is in a certain place in the city and want to take a bus, through its mobile device the user will know if the bus route is close to its location.

In this paper we present an application of AR based in *Android* oriented to SmartPhones that allow the user to enrich the physical information of the environment with virtual information such as routes of bus, lines that pass within 300 meters of the place in which the user is located, the arrival times, etc.

More specifically, the system can display the bus routes over the street that the user has in his front superimposed on the video stream on his mobile device. The user can also get information about a queried bus line, an estimate of arrival times for the next bus and a view of selected bus routes. The user's position is obtained from the *GPS*[3], orientation from the accelerometers and gyroscopes and georeferences data are downloaded from *OpenStreetMap* (OSM) servers.

The structure of this paper is as follows: The following section will provide an overview of the history of AR, AR on mobile devices and existing information systems relative to public transport. The third section will present the developed system architecture. Details of implementation will be provided on the fourth section. The fifth section will show the case study and finally outline the conclusions and future work are presented.

2 Background

We will introduce basic concept in references to AR, AR on mobile device and systems for visualization of maps and routes over them.

2.1 Augmented Reality and Mobile Devices

The term Augmented Reality (AR) is used to define a direct or indirect view of a real physical environment which elements are merged with virtual elements to create a real-time mixed reality. Guided by Figure 1 we can see where is placed the AR within the world of mixed reality [9].

In 1997 *Ronald Azuma* presented the first study of AR [4]. This publication established the physical characteristics of the AR, ergo the combination of real world and the virtual, real-time interactions and sensing in 3D.

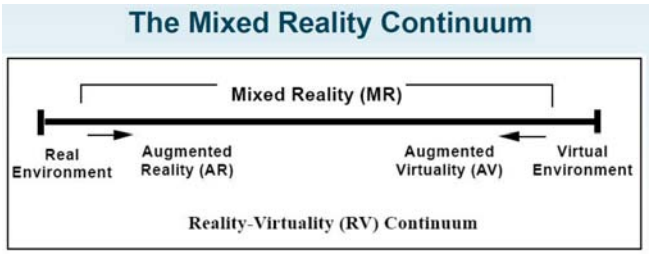


Fig. 1. Graphic illustration of the concept of Mixed Reality.

AR applications can be classified into two types: *indoor* and *outdoor*. While the former are used in closed environments and their goal is to work without

[3] Global Positioning System

user restrictions ([3],[6]), the latter are applications that have no environment restrictions.

Outdoor applications are based on two types of technologies: portable and immersive ([3],[6]). The first type consists in making a computer graphics overlapping on camera view of the portable device. In the second type must have generally, a *Head-Mounted Display* HMD that allows overlaying the images directly into the user's view, thereby achieving high levels of immersion.

In 1968 *Ivan Sutherland* created the first mobile AR system [14], which consisted of two trackers for correct positioning of images, each one with 6 degrees of freedom, one was ultrasonic while the other was a mechanic.

Later, in 1992, *Tom Caudell* and *David Mizell* first used the term *AR* [5] to refer to the computer image overlay on reality. At that time, the HMD, was the only means envisaged for mobile AR applications.

In later years there were two important developments: these were the Tobias Höllerer ([7],[2]) and Mathias Möhring ([10]). The first allowed to the user explore the story of a tourist spot through mobile device pointing it to different parts of the same spot, while the second developed a 3D tracking system for mobile devices and the screen displays information associated with AR mode.

Recently, research in this area (AR) has focused on mobile devices. In early 2000, developed projects such as *Bat-Portal* [11], it was based on *Personal Digital Assistant* (PDA) and technology *Wireless*. The PDA was used as a client to capture and transmit video to a dedicated server which performed the image processing and proceeded to render and compose 3D objects. While initially the prototypes were based on a distributed strategy to delegate the graphics processing, the fast advancement in mobile phones allowed the development of applications that recognize markers in the environment. Subsequently, with the integration of new sensors on devices and growth in computing processing power, the field of AR applications for mobile devices grew exponentially [10] [12] [15].

In 2007, Klein and Murray [8] presented a robust system capable of tracking in real time, using a monocular camera in a small environment. In 2008 *Wikitude* AR browser [1] was launched, it combines GPS and compass data entries in the *Wikipedia*. Finally, in 2009 *White* introduced *SiteLens* [16], an system and set of techniques for supporting site visits by visualizing relevant virtual data directly in the context of the physical site.

2.2 Maps Visualization and Routes

There are several alternatives when it comes to display maps on mobile devices. One of them is the version of *GoogleMaps* oriented phones, the software creates the same experience for the user as a query from *GoogleMaps* web page. Another alternative is *Mobile Gmaps* (MGMaps), it is developed with technology *J2ME*, for maps obtaining the system consults sources such as *Yahoo! Maps*, *Windows Live Local* [4], *Ask.com* and *Open Street Map* (OSM), the features are similar to those of *GoogleMaps*.

[4] MSN Virtual Earth

There are also applications that use the voice as an alternative to navigation; an application of this style is *Aura Navigation*, which allows to the user to view route maps and move respects these using the user's voice as a guide.

Finally, some applications that use AR for display are *Wikitude Drive* and *GPS Cyber Navi*. They show a route previously configured by the user, allowing it to reach its destination in a different way.

From the literature it can be seen that there is no previous works using AR oriented to bus routes displaying on mobile devices.

3 BusWay-AR

When users are in a particular bus stop, usually they can access to the bus line that arrive to the stop; generally, there is no information about arrival times, bus routes, their location, etc. Additionally, it is useful to know what are the bus lines that circulate in a certain radius close to them and where they circulate. This motivated us to develop an application that would provide information associated with buses lines that are in a user environment.

The application developed through AR interface provides the user who is located in a certain geographical position information about: accessed bus line, the route of this and arrival times. The system can track the user and display a 2D augmentation of bus routes showing the path, if the path is round trip or return route, in addition to other information. All of this information is added to the field of view on the mobile device video stream. The information of the bus lines is obtained from the *OSM* servers. The application was developed on a *SmartPhone* equipped with camera, 3G connectivity, Wirelees, GPS, accelerometers and gyroscopes.

The system can determine the position and orientation of the user, to obtain information concerning to the bus line, routes, arrival and departure, calculate the estimated arrival times to the user's position and finally, display all this information in a graphical interface, which overlaps layers of information to the video stream of the mobile device.

3.1 System Architecture

The proposed system consists of five sub-systems: the processing of the route, the position for obtaining and calculating arrival times, the rendering, the user interface and the AR.(Figure 2)

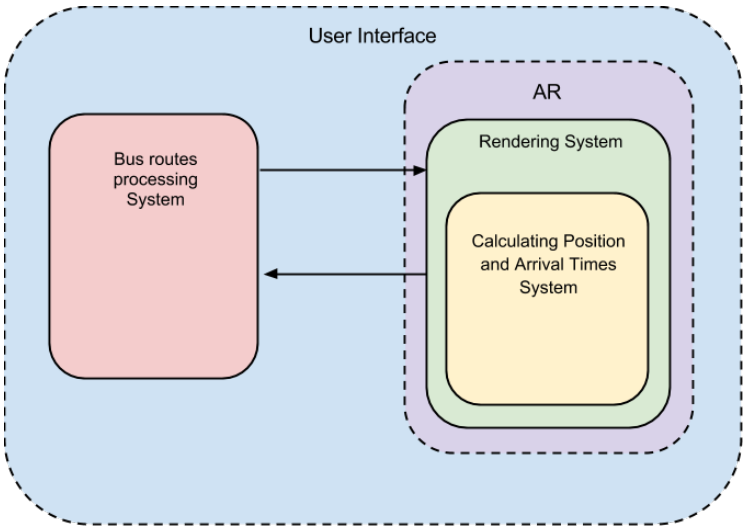


Fig. 2. System Architecture.

The *Route Processing Subsystem* is on charge of processing the path of the line selected by the user, which is communicated to the system via data provided by the graphical interface. Once the line has been obtained on request, is queried to *OSM* servers, the obtained information is stored in a suitable structure and is communicated to the *Rendering Subsystem*.

Obtaining Subsystem Calculating Position and Arrival Times is responsible for obtaining the user geographical position and the start time of the bus line consulted. From both, the subsystem proceed to calculate the arrival time and retrospective communicates to the *Rendering Subsystem*.

The *Rendering Subsystem*, which is responsible for generating the image displayed to the user, receives as input the path of the selected bus line and the user geographical position; it performs calculations for correctly positioning points of the route on the screen and also the bus position, it will be drawn only if the bus is within the range of view of the user. It is also responsible for providing information to the user, about arrival times, user geographical position, GPS status, etcetera.

The *AR Subsystem* is on charge of linking information from the *Rendering Subsystem* and the *Obtaining Subsystem Calculating Position and Arrival Times*, to be used for the system.

Finally the *User Interface* is the way by which the system communicates with the user, the latter being able to denote their needs and see the answers.

4 System Implementation

The system implementation was developed on *Android* and is intended to operate with bus routes of the city of *Bahía Blanca*.

The *Processing Subsystem* starts to work after the user selecting a bus line, the bus line is communicated to the subsystem that is responsible for making a query to the OSM page[5] where the bus line route is stored. The results are stored in an *XML* file.

For the development of *Positioning Subsystem*, we proceeded to obtain the user geographical position. *Android* provides several options, one of them is to use the built-in *GPS* sensor on the mobile device, to use it *Android* provides the *LocationProviders* data type, which can give us the position in two different ways: *GPS-Provider* and *Network-Provider*. We opted for the use of *GPS-Provider* as its accuracy was better.

Finally, the *Subsystem of Calculating Arrival Times* is responsible for telling the user how long it would take the next bus to arrive to the bus stop or where the bus is located, that was done by algorithms that estimate arrival times, based on data provided by the municipality of *Bahia Blanca* [6].

The *Rendering Subsystem* is responsible for drawing the scene viewed by the user. To this propose it should be taken into account the device orientation and the user interaction with the information displayed on the screen. Due to the complexity involved in the tasks outlined in the preceding paragraphs, we will see how their implementation were carried out.

For obtaining the image from the camera, *Android* provides access to the camera frames by modifying the main configuration file. Once we get the camera preview, we had to get the device orientation. *Android* provides us with a set of sensors, in particular, the sensor *TYPE_ROTATION_VECTOR* gives information about accelerometer and magnetic field. In this way we obtain the orientation of the device relative to the axis of the earth (aligned to the north), which is essential given that our system will use real positions (latitudes and longitudes).

To perform rendering of objects in the AR system, we used *OpenGL*, this gives us an API[7] with primitive graphics for drawing simple shapes. In our case, *Android* provides a special version of *OpenGL* for mobile devices, *OpenGL ES*. The version used was 1.0. Since we wanted to draw on the camera frames, we had to create a scene in our *OpenGL* space. The scene consists in objects, routes, which are in the *OpenGL* world space and an associated camera which will be rotated and moved in order to observe different objects from different points of view.

Since our main goal is to draw the routes of the bus line on the camera frame and the route consist on a set of latitudes and longitudes, we must convert those latitudes and longitudes from the world coordinate system to *OpenGL*

[5] http://wiki.openstreetmap.org/wiki/Bahía_Blanca/transporte_publico
[6] http://www.bahiablanca.gov.ar/conduce/transporte1.php
[7] Application Programming Interface

coordinate system. In order to perform the conversion, we had to keep in mind that we are referring to a geodetic coordinate system (latitudes and longitudes) and a geocentric coordinate system (*OpenGL*), in this way we have to made the respective transformations based on data provided by the next equations:

$$1/f = \text{flattenig factor} \tag{1}$$

$$e^2 = (a^2 \text{ - } b^2)/a^2 = 2f \text{ - } f^2 \tag{2}$$

$$v = \frac{a}{\sqrt{1 \text{ - } e^2 \sin^2 \varphi}} \tag{3}$$

$$X = (v + h) \cos \varphi \cos \lambda \tag{4}$$

$$Y = v + h \cos \varphi \sin \lambda \tag{5}$$

$$Z = [(1 \text{ - } e^2)v + h] \sin \varphi \tag{6}$$

where h is the *GPS* height and the variables a and b are the length of semi-major axis and the semi-minor axis of Earth respectively. λ is the longitude and φ is the latitude.

From above equations (based in [13] and [17]) were transformed latitude and longitude to X, Y and Z coordinates to draw the scene in *OpenGL*.

Needless to say, we used a *float* value in the internal representation for position, latitudes and longitudes are expressed in *double*; the systems perform the transformation with a lost of accuracy, therefore it is possible that in some cases there is shifting between the actual data and those generated by *OpenGL*.

5 System Testing

For testing we proceeded to the selection of the bus line 503 of the city *Bahía Blanca*. Since we want to analyse the system response under different circumstances, a prototype interface was developed, the interface can select the bus line manually, also the user must specify if he/she is in a default position or if the position can be get by the *GPS* (Figure 3).

Once the user have selected the option to show bus route, the route was displayed, both round trip (green) and return route (red), the GPS status (ON/OFF), latitude and longitude of the user geographical position, address (street/number), the arrival times from the round trip bus and the return route bus, this can be seen in Figure 4.

Data were obtained and the system was tested with both the *GPS* turned on and off, this can be seen in Figure 4.

It can be seen on the left of Figure 4 a map with way points recorded within a certain radius, you see green dots (round-trip), red dots (return journey) and a yellow dot(user's position). The right side of Figure 4 shows the view that the user has on the mobile device, the shifting between the bus route and the street is due to the *GPS* error and the rounding data error (move from *double* to *float*).

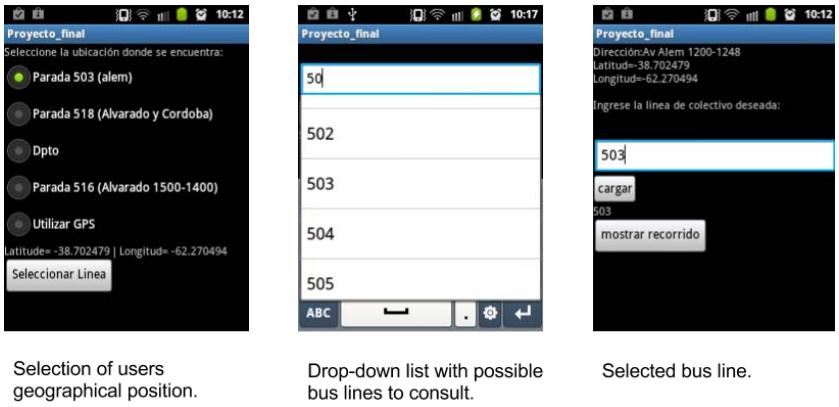


Fig. 3. Prototype Interface.

Fig. 4. Caso de test del recorrido línea 503.

6 Conclusions and Future Work

Despite the technological advances in recent years, there are still difficulties, for example, in obtaining the position and orientation in large areas, the size of displays and graphics processing capabilities. However, the AR is a useful and versatile alternative to organize and contextualize the information. We have presented the design and implementation of a 2D mobile AR application where the visualization of the route of a particular bus line is superimposed on the mobile device video stream with addition of estimated arrival times, get the user's position and display all this information contextualized in the user interface.

In the testing performed we highlight the problems generated by both *GPS* accuracy and the loss of precision due to transform latitude and longitude coordinates to *OpenGL* coordinates. These problems lead to a shift in the routes

visualization. Even though the objective of obtaining different kind of information about a particular bus line, these problems must be solved yet. About the positioning, it should work better with a higher precision *GPS* or with *DGPS*[8]. With respect to coordinate transformation, we should find an alternative representation in fixed point for latitude and longitude and with a defined range, perform a more accurate conversion into the *OpenGL* coordinate system.

In addition to seeking to solve the above problems, the future work is to be conducted online recognition of *OCR*, use version 2.0 of *OpenGL ES* and make the system has a 100% coverage information about bus lines. This paper is a starting point for the development of outdoor applications as we believe that this is a field of application in which mobile devices can be a very versatile alternative they record graphics on outdoor environments freely.

7 Acknowledgment

This work was partially funded by the project 24/N028 of Secretaría General de Ciencia y Tecnología, Universidad Nacional del Sur, PICT 2010 2657, FSTICS 001 "TEAC" and PAE 37079.

References

1. http://www.wikitude.com
2. Third International Symposium on Wearable Computers (ISWC 1999), San Francisco, California, USA, 18-19 October 1999, Proceedings. IEEE Computer Society (1999)
3. Avery, B., Smith, R.T., Piekarski, W., Thomas, B.H.: Designing outdoor mixed reality hardware systems. In: The Engineering of Mixed Reality Systems, pp. 211–231 (2010)
4. Azuma, R.T.: A survey of augmented reality. Presence: Teleoperators and Virtual Environments 6(4), 355–385 (Aug 1997)
5. Caudell, T.P., Mizell, D.W.: Augmented reality: an application of heads-up display technology to manual manufacturing processes. Proceedings of the TwentyFifth Hawaii International Conference on System Sciences 2, 659–669 (1992), `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=183317`
6. Gotow, J.B., Zienkiewicz, K., White, J., Schmidt, D.C.: Addressing challenges with augmented reality applications on smartphones. In: MOBILWARE. pp. 129–143 (2010)
7. Höllerer, T., Pavlik, J.V., Feiner, S.: Situated documentaries: Embedding multimedia presentations in the real world. In: ISWC [2], pp. 79–86
8. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. pp. 1–10. ISMAR '07, IEEE Computer Society, Washington, DC, USA (2007), `http://dx.doi.org/10.1109/ISMAR.2007.4538852`
9. Milgram, P., Takemura, H., Utsumi, A., Kishino, F.: Augmented reality: A class of displays on the reality-virtuality continuum. pp. 282–292 (1994)

[8] Diferenttial GPS

10. Möhring, M., Lessig, C., Bimber, O.: Video see-through ar on consumer cell-phones. In: Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality. pp. 252–253. ISMAR '04, IEEE Computer Society, Washington, DC, USA (2004), http://dx.doi.org/10.1109/ISMAR.2004.63
11. Newman, J., Ingram, D., Hopper, A.: Augmented reality in a wide area sentient environment. In: Augmented Reality, 2001. Proceedings. IEEE and ACM International Symposium on. pp. 77–86 (2001)
12. Newman, J., Schall, G., Barakonyi, I., Schürzinger, A., Schmalstieg, D.: Wide area tracking tools for augmented reality. In: In Advances in Pervasive Computing 2006, Vol. 207, Austrian Computer Society (2006)
13. OGP: Coordinate conversions and transformations including formulas. OGP Publication 373-7-2 – Geomatics Guidance Note number 7 2, 131 (2012)
14. Sutherland, I.E.: A head-mounted three dimensional display. In: Proceedings of the December 9-11, 1968, fall joint computer conference, part I. pp. 757–764. AFIPS '68 (Fall, part I), ACM, New York, NY, USA (1968), http://doi.acm.org/10.1145/1476589.1476686
15. Wagner, D., Schmalstieg, D.: First steps towards handheld augmented reality. pp. 127–135 (2003)
16. White, S., Feiner, S.: Sitelens: situated visualization techniques for urban site visits. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1117–1120. CHI '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1518701.1518871
17. Wikipedia: Geodetic system - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Geodetic_system (Agosto 2012)

Aplicación turística para dispositivos móviles basada en técnicas de visión computacional

Pablo A. Sosa†, Enrique M. Albornoz†‡ y César E. Martínez†§

† Centro de Inv. en Señales, Sistemas e Inteligencia Computacional (SINC(i))
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral
CC217, Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina
‡ CONICET, Argentina
§Laboratorio de Cibernética, Universidad Nacional de Entre Ríos
{psosa@quboo.com.ar,emalbornoz@fich.unl.edu.ar,cmartinez@fich.unl.edu.ar}

Resumen En la actualidad, los dispositivos móviles se han convertido en una herramienta útil que brinda al usuario una amplia variedad de funcionalidades. Además, la evolución tecnológica de éstos ha permitido que puedan ejecutarse algoritmos muy exigentes como aquellos utilizados en procesamiento de imágenes, reconocimiento de patrones, etc. Este contexto ha impulsado un nuevo campo de investigación y desarrollo, mientras que genera un nicho de mercado que apenas ha comenzado a explotarse a nivel regional. En este trabajo se presenta el desarrollo de una aplicación de información turística para dispositivos móviles que incorpora técnicas de visión computacional. Se trabajó sobre el circuito turístico del "Camino de la Constitución" de la ciudad de Santa Fe. El sistema realiza el reconocimiento automático de imágenes de los diferentes mojones informativos y provee al usuario una vasta cantidad de información.

Palabras claves: Turismo, OpenCV, Android, OCR.

1. Introducción

En los últimos años, se ha vuelto masiva la utilización de dispositivos móviles inteligentes. Su evolución tecnológica es tal que su capacidad de cómputo prácticamente iguala a las computadoras de escritorio, permitiendo ejecutar algoritmos cada vez más exigentes como los utilizados en procesamiento de imágenes, visión computacional, reconocimiento de patrones, entre otros [1]. La comunidad científica ha acompañado esta evolución generando un nuevo campo de investigación y desarrollo, migrando sus aplicaciones desde las computadoras de escritorio e inclusive diseñandolas para ser utilizadas dispositivos móviles como Tablets o Smartphones [2]. No obstante, el avance a nivel académico de este área no se corresponde con el mismo a nivel comercial, donde las aplicaciones usualmente son más básicas y no se ha logrado explotar el consumo masivo por parte de la población. Existen aplicaciones comerciales que realizan procesamiento de imágenes y a partir de la identificación de objetos de interés retornan información de utilidad. Entre las más conocidas se pueden citar *Google Goggles*, *Wikitude* y *Layar*.

Sin embargo, las dos últimas realizan este trabajo mediante la utilización del *sistema de posicionamiento global* (GPS: del inglés Global Positioning System.) del dispositivo móvil [3]. Google Goggles es un servicio de Google disponible para Android e iPhone que permite reconocer objetos mediante fotos capturadas con un dispositivo móvil y devolver resultados de búsqueda e información relacionada. Este sistema reconoce lugares, monumentos y textos, entre otras cosas. Su funcionamiento consiste en apuntar con la cámara del dispositivo móvil a un lugar conocido, un producto, un código de barras o un código de respuesta rápida (QR: del inglés quick response code). Luego, si Google lo encuentra dentro de su base de datos, ofrece información relacionada [4]. Wikitude es un software de realidad aumentada (RA)[1] para dispositivos móviles que fue desarrollado por la compañía austríaca *GmbH* y publicado en octubre de 2008 como software gratuito. Para obtener la localización de los objetos en la RA se utiliza la posición del usuario a través del GPS y la dirección en la que el usuario mira mediante el uso de la brújula y el acelerómetro. Esta aplicación depende de la precisión del GPS, que a veces no es lo suficientemente exacto [5]. Layar es un software que utiliza el GPS y la brújula de los dispositivos Android para ubicar la posición del usuario y su orientación. La cámara del dispositivo captura el entorno y reproduce la imagen en la pantalla, mientras que el software adiciona información de lugares tales como cafeterías, restaurantes, cines, etc. [6].

En este trabajo se presenta el diseño y desarrollo de una aplicación para dispositivos móviles con sistema operativo Android [7] basada en técnicas de visión computacional. La aplicación pretende ser una herramienta útil para los visitantes de la ciudad de Santa Fe, que visiten el circuito turístico del "Camino de la Constitución". Este es un recorrido museológico que integra 18 sitios y edificios de valor simbólico y arquitectónico, con el objetivo de reconstruir y narrar la historia de la relación de la ciudad de Santa Fe con la Constitución Nacional. La propuesta integra diversas funciones: histórica, cultural, educativa y turística, y pone de relieve nuestra vida política pasada y presente relacionada con los diferentes procesos y acontecimientos vinculados con la Carta Magna[2].

Una de las consignas fundamentales de este trabajo fue la de utilizar los mojones tal y como son actualmente, sin modificar su estética, emplazamiento, iluminación, etc. Entonces, en un primer paso se definió y realizó un importante relevamiento fotográfico utilizando diversos dispositivos móviles y condiciones de iluminación naturales, para generar una base de datos de imágenes de los mojones informativos. Luego se exploraron técnicas de preprocesamiento, análisis y clasificación de imágenes, orientadas a obtener el mejor desempeño manteniendo la simplicidad y una carga computacional baja, a fin de lograr una aplicación veloz. Finalmente, se diseñó una interfaz gráfica que permite al usuario tomar una fotografía del mojón (Figura 1) y obtener información turística relacionada.

[1] RA es el término que se usa para definir una visión directa o indirecta del mundo real, cuyos elementos se combinan con elementos virtuales para la creación de una realidad mixta en tiempo real.

[2] Más información en http://santafeciudad.gov.ar/cunadelaconstitucion/

Figura 1: Mojón informativo. (Imagen tomada de http://santafeciudad.gov.ar)

El resto del trabajo se organiza como se detalla a continuación. En la Sección 2 se describen los materiales y métodos utilizados. Mientras que la Sección 3 presenta el método propuesto considerando diferentes técnicas. Finalmente, en la Sección 4 se encuentran los resultados y conclusiones.

2. Materiales y tecnologías empleadas

El primer paso fue diseñar un protocolo para tomar las fotografías y generar una base de datos de imágenes. Se tomó un gran número de fotografías de los 18 mojones del Camino de la Constitución de la ciudad de Santa Fe, considerando siempre la mayor naturalidad posible en la toma para poder lograr el sistema más robusto posible. Se consideraron las siguientes condiciones:

1. Rotación: las fotos se capturaron con diferentes rotación hacia la izquierda o derecha.
2. Orientación del dispositivo: el dispositivo estuvo ubicado en forma vertical u horizontal, indistintamente.
3. Independencia del dispositivo: las capturas se realizaron para distintos dispositivos móviles:
   a) Smartphone Samsung Galaxy i5550, sistema operativo Android 2.3.
   b) Tablet ASUS Transformer TF101.CPU NVIDIA Tegra 2 1.0GHz. Android 3.2 Honeycomb O.S.
4. Múltiples resoluciones: las fotografías fueron tomadas con 1.3, 2.3 y 5 Megapíxeles.
5. Condiciones de iluminación: se consideraron condiciones naturales sin iluminación artificial y se tomaron de mañana, de tarde y de noche.

La base de datos conformada consiste en un conjunto de 693 fotografías. Éstas se utilizaron tanto para el preprocesamiento como para la clasificación de las imágenes.

La aplicación se desarrolló en Android para la versión 2.3 o superior, que posee compatibilidad con la biblioteca OpenCV[3]. Más específicamente, se utilizó Java [8] para desarrollar la lógica de la aplicación y XML para la interfaz de usuario, utilizando el entorno Eclipse [9]. Con respecto al desarrollo de rutinas para el procesamiento de imágenes se decidió utilizar la biblioteca OpenCV (versión 2.4.2) que es Open Source y multiplataforma [10]. Ésta provee un conjunto de funciones para procesamiento y análisis de imágenes, permite trabajar con video y tiene una versión disponible para Android.

3. Método propuesto

En esta sección se describe primeramente el preprocesamiento de las imágenes y la extracción de características, y luego se presenta la evaluación de varias alternativas para la clasificación de las imágenes. Finalmente, se presenta el desarrollo de la interfaz gráfica con la que el usuario interactúa.

3.1. Preprocesamiento de la imagen

En esta etapa podemos definir dos fases, en la primera se busca normalizar las imágenes tomadas desde el dispositivo móvil y en la segunda se realiza la extracción de la información distintiva del mojón.

Normalización de imágenes

Luego de evaluar varias alternativas, se obtuvo una secuencia de operaciones que permite preprocesar cualquier imagen para normalizarla:

1. Escalar la imagen: considerando el compromiso entre la velocidad de cómputo y los detalles en la imagen, el tamaño elegido es 800x600 píxeles.
2. Convertir a escala de grises: se realiza porque los procesos posteriores no precisan información del color y así se manipula menos información.
3. Ecualizar el histograma: se realiza para mejorar el contraste. Redistribuye de la forma más uniforme posible los grises de la imagen original sobre el total de intensidades [11].
4. Binarizar: es útil para la posterior detección de bordes.
5. Detectar bordes: se obtienen los bordes de la imagen aplicando el método de Canny [12]. Con esta técnica se obtuvieron mejores resultados que con el detector de bordes de Sobel [11].
6. Encontrar líneas principales: se utiliza la *transformada de Hough* que considera las relaciones globales entre píxeles permitiendo encontrar ciertos patrones en la imagen como líneas y círculos [13]. Aquí se determina el ángulo de rotación de la imagen capturada utilizando características particulares (marcos de los mojones) presentes en todas las imágenes.
7. Rotar: se rota la imagen utilizando el ángulo obtenido previamente.

En la Fig. 2 se puede ver el resultado de aplicar los pasos de la normalización.

[3] http://opencv.org/platforms/android.html.

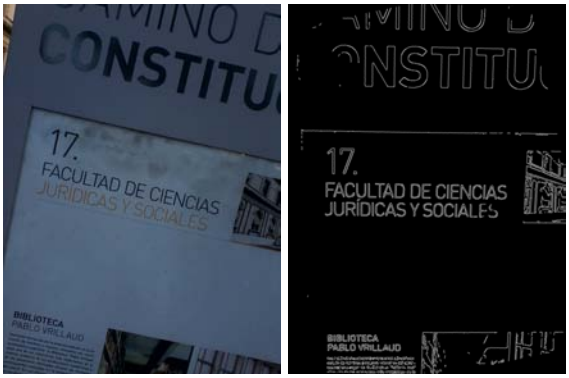


Figura 2: Imagen original y normalizada.

(a) Recorte de la zona de interés.

(b) Imagen de sólo números.

(c) Imágenes con los dígitos recortados.

Figura 3: Imágenes de la extracción del identificador.

Extracción del identificador único

Luego de realizar la normalización, se procede a extraer el número que fue el patrón escogido como identificador único para cada mojón. Para ésto se realiza el siguiente procedimiento:

1. Localizar: se extrae el cuadrante superior izquierdo de la imagen.
2. Dilatar: se realiza un proceso de dilatación con una máscara de 2x2, para que las líneas de los marcos de los mojones queden bien definidas, y así, facilitar el recorte del identificador.
3. Recortar zona de interés: se buscan las líneas horizontal y vertical correspondientes a los marcos de los mojones para hacer el recorte, ya que éstas son características que siempre están presentes en la imagen. En primer lugar se utilizan acumuladores de píxels no nulos a lo largo de las filas para encontrar la línea más extensa, luego se aplica el mismo proceso sobre las columnas. Es importante el orden de estas operaciones para salvar casos de mojones con perspectivas. El resultado de esta etapa se ve en la Figura 3a.
4. Erosionar: se realiza una erosión para que los dígitos esten bien separados.

5. Extraer número: se utiliza el acumulador de píxeles no nulos sobre filas para obtener la imagen del número basado en los espacios anterior y posterior (Fig. 3b).
6. Recortar los dígitos: se realiza utilizando el método de componentes conectadas [11]. Se seleccionan las dos componentes más grandes y se evalua la relación de aspecto de cada una para descartar ruido. En la Figura 3c se observa un ejemplo del resultado final.

3.2. Extracción de características

En esta etapa se extraen características relevantes de las imágenes de los dígitos. A continuación se presentan las alternativas consideradas.

Momentos invariantes de Hu

Los momentos invariantes de Hu representan una serie de características de los objetos independientemente de su posición, escala o rotación [11]. Hu obtuvo sus invariantes a través de los momentos geométricos. Éstos se definen como :

$$u_{pq} = (x - \hat{x})^p (y - \hat{y})^q f(x, y) dx dy \tag{1}$$

donde u_{pq} es el momento geométrico de orden $(p + q)$, $f(x, y)$ es el valor del píxel en la posición (x, y) y $(\hat{x}, \hat{y})$ son las componentes del centroide. Partiendo de los momentos definidos anteriormente, se definen los coeficientes n_{pq} que son invariantes al escalamiento y su expresión viene dada por:

$$n_{pq} = \frac{u_{pq}}{u_{00}^{1+\frac{p+q}{2}}} \tag{2}$$

Utilizando n_{pq} se obtienen las expresiones matemáticas que dan lugar a los siete momentos invariantes de Hu.

Vector de distancias al primer píxel no nulo de una imagen.

Este método utiliza la imagen binaria del dígito obtenida previamente. Se crea un vector columna de longintud igual al alto de la imagen cuyos valores, para cada fila, representan las distancias al primer píxel no nulo de la imagen, en el sentido izquierda a derecha. Luego, se crea otro vector considerando las distancias de derecha a izquierda. Esta información es representativa del contorno del número en la imagen[4].

[4] Para esta etapa, las imágenes están normalizadas en 125x65 píxeles.

3.3. Métodos de Clasificación

A partir de imágenes limpias y manualmente ajustadas se calcularon los patrones de referencia para los momentos invariantes de Hu y vectores de distancias no nulos. El paso siguiente es comparar las características calculadas para una nueva captura con las características de los patrones de referencia. Para realizar ésto, y considerando la velocidad de cómputo, se decidió clasificar utilizando una medida de distancia basada en la *norma-2*, así el patrón será clasificado considerando la menor distancia euclídea:

$$\text{Clase}_{\text{Y}} = \min_{j} \left\{ \sqrt{\sum_{i=1}^{n} (\mathbf{X}_i^j - Y_i)^2} \right\} \tag{3}$$

donde $\mathbf{X}^j$ son los vectores de referencia, Y el patrón a clasificar y n la dimensión de las características.

Método de clasificación alternativo

Como alternativa de clasificación se consideró la utilización del reconocimiento óptico de caracteres (OCR: del inglés Optical Character Recognition). Se investigaron varias alternativas de OCR y finalmente se seleccionó Tesseract OCR [14]. Ésta es una biblioteca libre con licencia *Apache License 2.0* escrita en C++, es muy eficiente y ampliamente utilizada[5].

Aquí también se realiza el preprocesamiento completo para obtener los dígitos, sin embargo, se recortan los dígitos a color para ser utilizados en el OCR. El uso de las imágenes de sólo dígitos mejora el rendimiento del OCR.

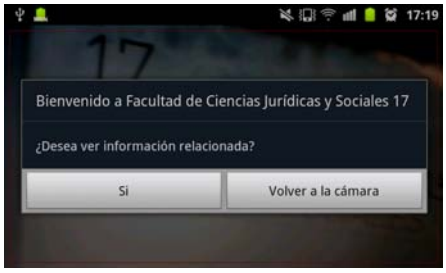
3.4. Interfaz de usuario

Una consideración general es que la información multimedia que provee la aplicación no se almacena en el dispositivo, a fin de utilizar la menor cantidad de recursos posibles, sino que se obtiene directamente desde la web. Por lo tanto, se requiere acceso a Internet mediante 3G o Wi-Fi para poder visualizar las imágenes, videos mediante streaming y enlaces a páginas web. El diseño de la interfaz de usuario es simple y permite rápidamente acceder a la información relacionada. La pantalla inicial permite acceder al sistema o encontrar ayuda acerca de cómo utilizar el sistema (Figura 4). El sistema solicita al usuario que tome una fotografía del mojón (Fig. 5a) e inmediatamente después de realizar el procesamiento y clasificación, arroja el resultado del mojón reconocido (Fig. 5b). En una etapa posterior el sistema brinda información en diversas alternativas multimedias (Fig. 6).

[5] Disponible en https://code.google.com/p/tesseract-ocr/.

Figura 4: Pantalla inicial.

(a) Captura de imagen.



(b) Mojón reconocido.

Figura 5: Capturas del sistema.

Figura 6: Información multimedia relacionada.

4. Experimentos y resultados

A continuación se presentan algunos resultados preliminares sobre los diferentes métodos propuestos. Para este experimento se utilizaron 75 imágenes (aproximadamente 5 de cada mojón), capturadas con iluminación natural (de mañana y tarde) y con el dispositivo Samsung Galaxy en 1.3 mpx. Se obtuvo un 78, 6 % de segmentaciones correctas hasta la etapa de recorte de los dígitos. Luego, se evaluó el tiempo total de los tres métodos de clasificación, ya que el

Tabla 1: Resultados de clasificación.

Método de clasificación	Aciertos	Tiempo de ejecución
Vector de distancias a píxeles no nulos	100 %	5475 milisegundos
Momentos invariantes de Hu	66.1 %	1329 milisegundos
Tesseract OCR	100 %	298 milisegundos

tiempo de preprocesamiento (hasta el recorte de los dígitos inclusive) es el mismo para todos los casos. También se registró la tasa de aciertos en el reconocimiento de los diferentes mojones. En la Tabla 1 se pueden ver algunos resultados preliminares de los tres métodos de clasificación. Estos resultados son promedios de evaluar los 59 casos en que la segmentación fue exitosa.

Los momentos invariantes de Hu mostraron cierta inestabilidad, a pesar de ser invariantes a la rotación y escalado y los resultados obtenidos son inferiores a los esperados. El método basado en las distancias demostró ser muy bueno a pesar de su sencillez, aunque requiere mucho tiempo de cómputo. Finalmente, el OCR demostró ser el que obtuvo más aciertos y respecto de los tiempos de ejecución, también Tesseract OCR fue le de mejor rendimiento. Evidentemente, la opción de clasificación elegida para la aplicación final es Tesseract por su desempeño y velocidad de cómputo.

5. Conclusiones y trabajos futuros

En este trabajo se ha presentado una aplicación turística para sistemas Android basada en técnicas de visión computacional. La aplicación reconoce automáticamente las imágenes de los mojones del circuito turístico *Camino de la Constitución* de la ciudad de Santa Fe y devuelve información multimedial relacionada. Se ha realizado una base de datos de imágenes considerando diferentes condiciones de iluminación, resoluciones de la cámara y dispositivos de captura. Se han evaluado diferentes alternativas de procesamiento de imágenes para lograr un adecuado preprocesamiento de las fotografías obtenidas por el usuario y se evaluaron diferentes clasificadores. Se diseñó una interfaz simple e intuitiva para la aplicación. Finalmente, se ensambló el sistema considerando los mejores resultados obtenidos en la etapa de desarrollo-evaluación para lograr una aplicación simple, potente y veloz.

Como trabajo futuro se propone mejorar el método de preprocesamiento para lograr una mayor tasa de segmentaciones correctas. También se considera hacer una evaluación más profunda del comportamiento del sistema considerando otras condiciones en la captura la fotografía, por ejemplo con imágenes nocturnas.

Agradecimientos

Los autores desean agradecer a la *ANPCyT* y *Universidad Nacional de Litoral* (proyectos PAE 37122, PACT 2011 #58, CAI+D 2011 #58-511) y al *CONICET*, por su apoyo.

Referencias

1. Daniel Wagner, Gerhard Reitmayr, Alessandro Mulloni, Tom Drummond, and Dieter Schmalstieg. Pose tracking from natural features on mobile phones. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR '08, pages 125–134, Washington, DC, USA, 2008. IEEE Computer Society.
2. Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21(6):34–47, 2001.
3. Zornitza Yovcheva, Dimitrios Buhalis, and Christos Gatzidis. Overview of smartphone augmented reality applications for tourism. *e-Review of Tourism Research*, 10(2):63–66, 2012.
4. Google Goggles. `http://www.google.com/mobile/goggles/`. Último acceso: Mayo-2013.
5. Wikitude. `http://www.wikitude.com/en/`. Último acceso: Mayo-2013.
6. Layar - Augmented Reality Browser. `http://www.layar.com/`. Último acceso: Mayo-2013.
7. Nisarg Gandhewar and Rahila Sheikh. Google Android: An emerging software platform for mobile devices. *International Journal on Computer Science and Engineering*, 1(1):12–17, 2010.
8. Javier Garca de Jaln de la Fuente. *Aprenda Java como si estuviera en primero*. Aprenda ..., como si estuviera en primero. Universidad de Navarra. Escuela Superior de Ingenieros Industriales, 1999.
9. Jeff McAffer, Jean-Michel Lemieux, and Chris Aniszczyk. *Eclipse Rich Client Platform*. Addison-Wesley Professional, 2nd edition, 2010.
10. Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, 2008.
11. Rafael Gonzalez and Richard Woods. *Digital Image Processing (2nd Edition)*. Prentice Hall, 2002.
12. John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.
13. Richard Duda and Peter Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.
14. Ray Smith. An overview of the Tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

A case study on 3D Virtual kitchen design with Kinect sensor

Matias N. Leone, Mariano M. Banquiero, AndresBursztyn

Proyecto de Investigación "Explotación de GPUs y Gráficos Por Computadora", GIGC - Grupo de Investigación de Gráficos por Computadora, Departamento de Ingeniería en Sistemas de Información, UTN-FRBA, Argentina

{mleone, mbanquiero}@frba.utn.edu.ar
andresb@sistemas.frba.utn.edu.ar

Abstract. The industry of kitchen design and furniture fabrication has been largely aided by 3D virtual scene applications and real-time visualization software. Microsoft Kinect provides skeleton tracking and bone orientation features than can be exploited in order to improve the interaction between the user and the 3D virtual scene representation.In this work we present a case study of an interactive application that assists the kitchen design process utilizing Microsoft Kinect hardware. A 3D avatar mesh mimics the user movement in real time, gesture recognition techniques are applied to produce actions on the virtual scene and the user can interact with custom made UI widgets using the hands as mouse cursors.

Keywords: Skeleton tracking, Gesture recognition, Kitchen design, DirectX

Fig.1. Left: 3D kitchen scene rendered by the application. Middle: the user interacts with the software using his hands. Right: the avatar follows the movement of the user in real time.

1 Introduction

Lepton Systems [1] is a software company that provides a desktop application for kitchen design. It let the user to choose appliances from the majority of the industries providers, visualize them, interact with them and set their location in a 3D representation of the kitchen. The application comes with a real time rendering engine that allows the user to envision a virtual representation of the kitchen being designed. It also provides tools for material optimization, costs estimation, reports generation and Autocad [2] model importers.

The company was planning their stand for the industry bigger exhibition celebrated in the country, and they wanted to show a state of the art demonstration of their technological capabilities for kitchen design.

In order to achieve this objective they asked our help to develop a software solution for 3D virtual interaction (Fig. 1). The goal was to build an application that lets the user of the exhibition to interact with the kitchen designed using his own hands. Allowing the user to change the material of the appliances, modify handles and terminations, and navigate through all the corners of the kitchen. The request had the aim to provide a better degree of interaction and immersion than the traditional input approach of mouse and keyboard.

The application developed makes use of Microsoft Kinect [3], which is a hybrid hardware-software solution for 3D skeletal tracking. Its sensors capabilities, its processing power and the low cost of the hardware make it an adequate candidate to fulfill the request. The features of Kinect used in this project are 2D hands tracking for UI navigation and 3D skeletal tracking for gesture recognition and avatar controlling.

The software was designed with three main modes of interaction:

- UI interaction mode: the user controls two mouse pointers with his own hands and navigates in 2D through different UI widgets, like buttons, scrollbars and message boxes, that allow him to interact with the scene.
- Avatar mode: a human 3D mesh is rendered and updated based on the skeleton data tracked by Kinect. The user can move his body and the avatar mimics his movements in real time. The user can also do some simple gestures that are recognized by the application to produce specific actions on the scene, like pulling his hand to open a drawer or waving his hand to close a door.
- 3D navigation mode: the user can move his hands to navigate through the scene in three dimensions. The right hand controls the position of the camera and may also alter the orientation. This allows the user to visualize the kitchen scene from different points of view.

2 Related work

The entertainment and videogame industry has been long trying to find alternatives to the classical joystick input in order to increase the level of realism of their titles. Nintendo Wii [4] was one of the first consoles to introduce a joystick with motion sensing capabilities through the use of accelerometer and optical sensor technology.Many attempts have been made to takes the Wii control to general applications beyond videogames, like the works described in [5][6][7].

Sony also provides PlayStation Move [8] as an alternative for PlayStation 3 console joystick. A complete reference ofvideogame motion controls and its evolution over time can be found in [9].

The Kinect for Windows Human Interface Guidelines [10] has importantremarks that should be taken into account when a new Kinect application is developed. Basic steps to integrate Kinect in a Windows application are described in [11]. A detailed explanation of some internal Kinect algorithms is described in [12], and [13]

explained the smoothing techniques used to filter undesired noise of the hardware sensors. Advanced hands and finger recognitiontechniques can be found in [14] and [15].

The project Ludique's Kinect Bundle [16] allows the user to employ the Kinect camera to interact with Windows UI in many ways. And Kineticspace[17] is a generic tool for gesture tracking and recording using Kinect.

The work in [18] also explores the use of Kinect in a kitchen but not from a design perspective as this paper, but for making use of Kinect in real-life kitchens for a variety of interactions, especially when your hands are full or messy,hence touchableand others traditional interfaces are not an option.

3 Application overview

Three modes of interaction were designed for the software. Each mode allows different kind of actions and makes use of diverse features of Kinect SDK:

3.1 UI Interactionmode

In this mode the user can moves his hands to control two mouse cursors in the screen. In every frame, the application retrieves the 3D position of both handsfrom Kinect, and then converts them to 2D screen positions. Each hand has its own delimited area of interaction (Fig.2). The bounding of the screen are adapted for each hand convenience.

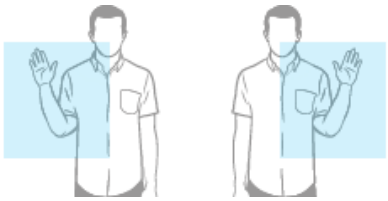


Fig.2. Screen bounding box of interaction computed for each hand.

For each hand, a screen bounding box is computed (physical interaction region) based on the distance between the hip center bone and the hand.Then the 3D position of each hand is converted to screen coordinates.

It should be noted that when Kinect cannot track the hand position, the previous screen position is used in order to avoid noisy movements.

The 2D coordinates of each hand is then used to place a cursor in the screen. The cursor movement allows the user to interact with different UI controls. The actions available are:

- To choose the kitchen scene to interact with from a list of pre-designed scenes.
- To change the material of all kitchen appliances, choosing between different types of surfaces, such as wood, metal and plastic.
- To change the handle of the doors and drawers of the kitchen.
- To change the mode of the application to avatar mode or 3D navigation mode.

The interaction with the UI was designed from scratch to be correctly adapted for Kinect movements. The buttons have a considerable screen size and are placed vertical aligned, in order to be comfortable to the user (Fig. 3).

Fig.3. Application UI widgets specially designed for comfortable Kinect interactions.

3.2 Avatar mode

In this mode the Kinect skeleton is tracked and used to render an avatar in the scene. The avatar is facing to the kitchen and gives his back to the camera. It is aligned with the 3D axis and it is watching towards negative Z.

The avatar is composed of many individual meshes that represent its body parts: arms, legs, torso, hands, feet, neck and head (Fig. 4). The 3D position and orientation of each bone is tracked and used to place the individual meshes correctly.

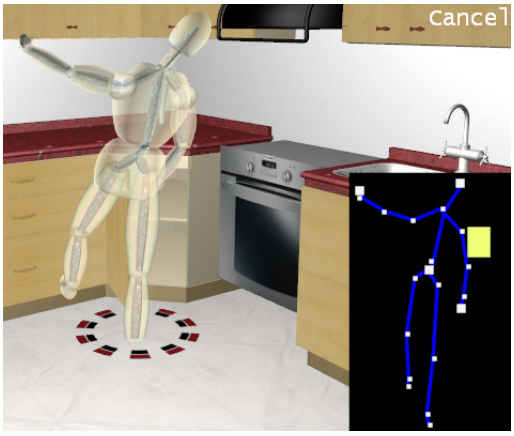


Fig.4. Left: avatar controlled by Kinect tracked skeleton. It is composed by individual meshes for each body part. Right: 2D tracking information of the skeleton.

Each individual mesh is initially located at the origin of coordinates pointing up to the positive Y axis. In every frame the mesh is rotated, translated and scaled according to its associated bone, using the following procedure:

- Restore the original position, orientation and scale of the mesh.
- Rotate the mesh using the orientation matrix tracked by Kinect for the start joint of the bone.

- Translate the mesh to the position of the start joint. This way the center of the mesh is located exactly where the join is, and its direction points to the end join of the bone.
- The mesh is scaled along its local Y axis according to the distance between the start joint and the end joint of the bone.
- To scale the XZ dimensions of the mesh, two approaches were tested. The first one consists of performing a scale proportional to the amount scaled in Y. The second approach is to have a fixed width and length for each mesh and not scale them at all. This second approach provides us with a better control of the avatar visual features.

This strategy allows the avatar to mimic the user movements in real time, which considerably increases the level of integration between the user and the application.

In order to allow the user to interact with the virtual kitchen being designed, the software is capableof detecting some simple hand gestures. The ones recognized by the application are:

- Move the hand forward and backward in a straight line: used to open or close drawers of the kitchen (Fig. 5).
- Wave the hand left to right or right to left: used to open or close doors of the scene.

In each frame,the application tracks both hands 3D positions and stores them in a buffer of one hundred previous frames. All these positions are treated as a cloud of points that need to be processed to extract some useful information. A statistical analysis is performed over this buffer in order to gather indicators that help to recognize the gesture. For each axis (x, y, z) the following indicators are computed: minimum value, maximum value, mean, variance and average of derivatives.

These measures are studied in each frame to detect valid avatar gestures. For example the open drawer gesture is triggered when:

- The average of derivatives in Z is positive
- The variance in X is almost cero
- The variance in Y is almost cero

This way the gestures are defined in a declarative fashion, based only on the values of this statistical indicators, and not on their absolute physical positions.

Once a gesture is detected the next step is to analyze if some object of the scene is close enough to be affected by the action. The application computes the screen distance between the average gesture position and the projected center of the object. The closest object is selected, provided the distance is below a given threshold, and the corresponding gesture action is rendered.

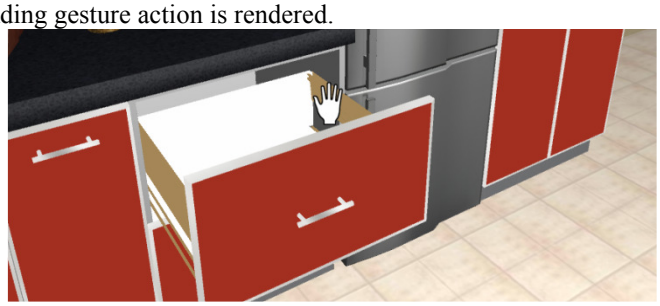


Fig.5. Opening drawer gesture recognized by the application.

Another approached was used to detect more complex gestures. First the required gestures are recorded in a custom made tracking application (Fig. 6). This tool allows the user to start and stop recording and then to trim the undesired parts the record. The tool tracks the movement of the hand for each frame and stores its 2D position and interval in a data file.

These gesture files are then used by the real time application to detect actions from the user and produce accordingly events in the kitchen scene. The application analyzes the last hands positionsof some frames and compares this buffer againstthe data of recorded gestures. The last one hundred frames are used as the size of the buffer being analyzed. Each value is compared against the recorded position of the gesture. If all distance of every tracked position and its corresponding recorded position is less than a given threshold then the application detects a gesture match.

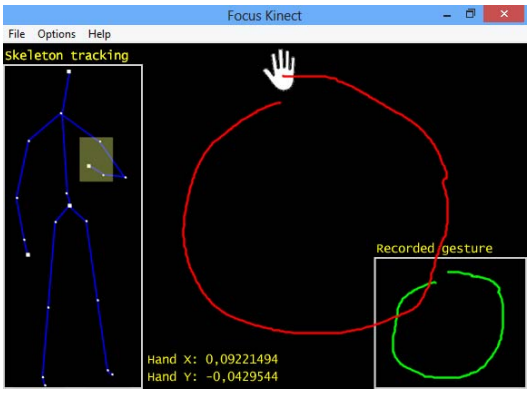


Fig.6. Tool developed for gesture recording.

But in order to accurately compare both gestures they must be previously transformed to some common frame. We callthis frame Gesture-space. First the samples are normalize to a space of [-1,1] x [-1,1], where the (0, 0) is the center of the gesture, which allow us to normalize the gesture scale. Then the first sample is positioned in (-1, 0) and the rest of the samples are rotated according to the angle of this first point. This step normalizes the gesture initial position and orientation. Finally the time of the gesture must be normalized, regenerating samples at regular intervals, which requiresadding or removing points so that both gestures may achieve the same amount of values. The total length traveled by the gesture is taken and divided by the samples count to obtain the segment size. The sequence is reconstructed based on this fixed segment.

Once the samples pass through this normalization steps both gesture can be compare as taking the distance of two functions. If this distance is lower than a given Epsilon, then both gestures can be considered equal.

This way the gesture detection procedure does not depends on absolute positions, speeds or any other physical properties, but on statistical measures, like average and standard deviation.

3.3 Navigationmode

In this mode the user can move around the kitchen scene using his right hand to control the camera. The movement is made in three dimensions. If the user moves up his right hand then the camera goes forward, and if he moves it down the camera goes backward. This axis controls the position of the camera, or look from. When he moves the hand left to right or right to left the user controls the orientation of the camera, or look at. This allows him to turn and watch different angles of the scene.

A map is shown at the right corner of the screen with a top view of the camera location and orientation in the scene.

4 Avatar tracking

The Kinect SDK sometimes cannot accurately capture all bones of the body, especially when one bone is occluded by another. For example this usually happens when the hand is exactlyin front of the shoulder. This kind of problems must be taken into account by the application in order to avoid undesired side effects, such as bad gesture recognition and unnatural avatar poses.

To avoid theseartifacts the software developed has an initial phase of skeleton calibration. During this phase, which last always the same fixed amount of time, all bones are tracked and stored in a buffer. Then some general measures of the skeleton are computed from this buffer, such as distance from hip to floor, length between the head and the hip, length of arms and legs, etc. These anthropomorphic features help the application to smooth and correct the data tracked from Kinect.

It is important that the calibration phase endure some seconds, in order to take an average of all these features. Otherwise incorrect initial measures may be used during all the application life-cycle.

For the Avatar mode the application takes the data tracked by Kinect but do not use it directly into their internal algorithm. A collection of human-constrain heuristics are applied in order to check that positions and orientations retrieved from the sensor areindeed valid for a human being. For example the length of the arms and legs must be equals to the original length captured in the calibration phase. The rotation matrix of knees, feet, elbows and neck cannot overpass allowed angles of motion for a human body. When these situations are encountered, the application ignores the current tracked value and reuses the last valid information for that bone. The same is applied when Kinect cannot track a bone at all.

All the coordinates captured by Kinect aregiven in “Kinect space”, but the Avatar mode needs to show the character interacting with the real dimensions of the kitchen. To achieve this, the application transformed the given coordinates from Kinect space to World space, using the following procedure:

- The hip position computed in the calibration phase is subtracted from all the other bones positions.
- The Z coordinate is inverted to achieve an avatar that gives its back to the user. This way the character can interact with the kitchen scene.
- The values are scaled to the scene measures.

After these steps, the individual meshes that compose the “avatar body” are placed in each bone with the procedure described in the previous sections.

5 UI

The user interface included in the application was specially designed to be adapted for Kinect style of movements and interactions. Widgets like buttons, scrollbars, list box and message box were developed using DirectX projected triangles with texture mapping.

First a low level GUI layer was built with many primitive instructions, such as draw line, draw rectangle, draw polygon, draw rounded rectangle, draw circle, draw disc, draw arc and draw image. All of them were developed with a combination of DirectX triangles fan, vertex color and texture sprites.

Then a higher level layer was built upon the last one to create the following UI widgets: button, displayable menu, frame, circle button, scroll button, message box, and progress bar. For example a button is render combining a draw image primitive with a draw rounded rectangle operation to show when the widget is selected. A frame is a floating panel that can contain other widgets within. A message box is created with a frame and two buttons for the "accept and cancel" options. The progress bar is rendered with many rounded rectangles and one main rectangle that vary its size according to the current progress.

All the buttons and menu items are selected with a timeout approach. The user controls two mouse cursors with his hands and when one of them is over a button a timer is started. If the user let the hand over this button for some time, then the button is selected. A circled progress bar is rendered around the cursor to give feedback about the elapsed time.

One problem with this approach is that the user tends to get his shoulder tired after some minutes of interaction, and then his hands begin to tremble. This flickering produces small movements on the screen and usually the hands get out of the selected button. When this happen, the widget timer is restarted and the task of accurately selecting one specific button becomes a frustrated action.

Our application solves this problem by reducing the amount of movement allowed to the cursor when it stays inside a button. When the hand is over a widget than can be selected, the application automatically centers the cursor to the middle of it (snap to grid). All small movements of the hand are discarded. If the user wants to exit the button then he has do an abrupt change in position. This strategy reduced the flickering and makes the UI easier to use.

6 Implementation and Results

We used Microsoft Kinect SDK 1.7 to track skeleton positions and orientations. These tracking features were integrated in our own real time rendering engine developed with C# 4.0 and Microsoft DirectX 9.

The kitchen scenes and furniture meshes were loaded from the Lepton systems proprietary format used in its software Focus 3D 2013 [1].

The application was run in a notebook with Windows 8, Intel Core i5 1.7 GHz with 8 GB Ram and Intel HD Graphics 4000 GPU.

The notebook also displayed the application through a projector located in the roof, so when one person was interacting with the software others were able to see what he was doing (Fig. 7).

Fig. 7. Top: a user of the exhibition selecting a material with his hands through the application. Bottom: other people watching the projected image of the software.

7 Conclusions and future work

A better degree of interaction between the user and the application waseffectively achieved and Lepton Systems was greatly satisfied with the results. The public at the exhibition was intrigued by the new experience of using his body to move around a virtual kitchen. Some vendors even propose to adapt the idea to other kind of business.

One of the main drawbacks we could observe during development is that Kinect cannot replace the traditional mouse and keyboard solution for many scenarios. Using your hands to move a screen cursor is not as precise as using the mouse. And when an application required a great degree of control the user may get frustrated with Kinect.

We believe that Kinect should be chosen for cases where traditional inputs do not apply. For example when the user cannot use his hands to touch the computer because he is doing something else or when all parts of the body may contribute to the action and not just the hands.

Another problem we encountered at the exhibition was that Kinect sometimes get confused when many people were near the range of the sensor. A common situation was when one user was interacting with the application and suddenly other person crossed behind. In most cases the software lost all the tracking and an application restart was required.

One feature of Kinect that could be integrated to this application is the use of speech recognition. The combination of hands gestures and spoken orders may increase the level of interaction.

Future improvements from the next version of the hardware (Kinect One [19]), specially about fingers tracking, may open many new interesting alternatives of interaction.

8 References

1. Lepton Sistemas SRL Soluciones inteligentes. 2013. Software for kitchen design and material optimization.
2. Autodesk. 2013. AutoCAD. 3D CAD Design Software.
3. Microsoft Corp. 2013. Kinect for Windows SDK.
4. Nintendo Co., Ltd. 2006. Nintendo Wii. Home video game console.
5. NithinSanthanam. 2012. Wii remote as a web navigation device for people with cerebral palsy. In Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '12). ACM, New York, NY, USA, 303-304.
6. David Scherfgen and Rainer Herpers. 2009. 3D tracking using multiple Nintendo Wii Remotes: a simple consumer hardware tracking approach. In Proceedings of the 2009 Conference on Future Play on @ GDC Canada (Future Play '09). ACM, New York, NY, USA, 31-32.
7. Sergio Albiol-Pérez, José-Antonio Gil-Gómez, Mariano Alcañiz, Roberto Llorens, and Carolina Colomer. 2012. Use of the Wii balance board system in vestibular rehabilitation. In Proceedings of the 13th International Conference on Interacción Persona-Ordenador (INTERACCION '12). ACM, New York, NY, USA, Article 11, 4 pages.
8. Sony Computer Entertainment. 2010. PlayStation Move. Motion-sensing game controller.
9. Sung, Kelvin. 2011. Recent Videogame Console Technologies. IEEE Computer 44.2: 91-93.
10. Microsoft Corp. 2013. Kinect for Windows Human Interface Guidelines v1.7.0. Microsoft. URL http://msdn.microsoft.com/en-us/library/jj663791.aspx.
11. Catuhe, David. 2012. Programming with the Kinect for Windows Software Development Kit. O'Reilly Media, Inc.
12. Shotton, Jamie, et al. 2013. Real-time human pose recognition in parts from single depth images. Communications of the ACM 56.1 (2013): 116-124.
13. MehranAzimi. 2012. Microsoft Advanced Technology Group. Skeletal Joint Smoothing White Paper.
14. Zhou Ren, JingjingMeng, Junsong Yuan, and Zhengyou Zhang. 2011. Robust hand gesture recognition with kinect sensor. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 759-760
15. Oikonomidis, Iason, NikolaosKyriazis, and Antonis A. Argyros. 2011. Efficient model-based 3D tracking of hand articulations using Kinect. BMVC.

16. Ludique's Kinect Bundle (LKB). 2012. Open source UI manager integrated with Kinect. URL https://code.google.com/p/lkb-kinect-bundle/
17. Kineticspace. Training, Analyzing and Recognizing 3D Gestures. 2011. URL https://code.google.com/p/kineticspace/
18. Galen Panger. 2012. Kinect in the kitchen: testing depth camera interactions in practical home environments. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '12). ACM, New York, NY, USA, 1985-1990.
19. Microsoft Corp. Xbox One. 2013. Successor video game console to the Xbox 360.

Vertex Discard Occlusion Culling

Leandro R. Barbagallo, Matias N. Leone, Rodrigo N. Garcia

Proyecto de Investigación "Explotación de GPUs y Gráficos Por Computadora", GIGC – Grupo de Investigación de Gráficos por Computadora, Departamento de Ingeniería en Sistemas de Información, UTN-FRBA, Argentina

{lbarbagallo, mleone, rgarcia}@frba.utn.edu.ar

Abstract. Performing visibility determination in densely occluded environments is essential to avoid rendering unnecessary objects and achieve high frame rates. In this work we present an implementation of the image space Occlusion Culling algorithm done completely in GPU, avoiding the latency introduced by returning the visibility results to the CPU. Our algorithm utilizes the GPU rendering power to construct the Occlusion Map and then performs the image space visibility test by splitting the region of the screen space occludees into parallelizable blocks. Our implementation is especially applicable for low-end graphics hardware and the visibility results are accessible by GPU shaders. It can be applied with excellent results in scenes where pixel shaders alter the depth values of the pixels, without interfering with hardware Early-Z culling methods. We demonstrate the benefits and show the results of this method in real-time densely occluded scenes.

Keywords: Occlusion Culling, Visibility Determination, GPU, Shaders

1 Introduction

Complex scenes with thousands of meshes and expensive shading computations are increasingly frequent in current real-time graphics applications. Although commodity hardware continues to increase its computational power every day, scenes like these cannot be directly supported at real-time frame rates. Optimization techniques are crucial in order to manage that kind of graphics complexity.

Frustum culling is a commonly used technique to avoid rendering meshes that are outside the viewing volume. These invisible models can be discarded at an early stage in the pipeline obviating expensive computation that will not contribute to the final image. Unfortunately it does not consider objects (occludees) that do not contribute to the final image because they are being blocked by others in front of them (occluders).

Because of this, several Occlusion Culling techniques were developed to overcome this limitation. Applications with expensive pixel shaders may greatly improve their performance by reducing fragments overdraw.

The Z-PrePass [1] technique avoids computing unnecessary pixel shaders following a two step procedure. First it draws the entire scene in order to store in the

Z-Buffer all the depth values of the scene visible points. Second the scene is drawn once more, but this time the GPU can early reject the occluded fragments based on already present depth values in the ZBuffer. This way non visible pixel shaders are not executed.

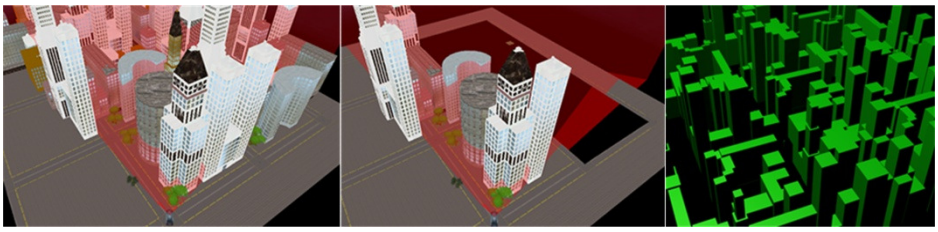


Fig. 1: Left: The densely occluded scene as viewed from the camera. Middle: The Occlusion Culling algorithm avoids rendering completely occluded objects. Right: The simplified occluder set used for occlusion.

This technique is used by many applications to reduce its pixel overdraw but its main limitation is that GPU cannot take advantage of the Early-Z [2] or [3] optimization when the pixel shader uses a depth writing operation [4], [5]. Since our method discards occluded objects before they get rasterized, no restrictions related to depth writing are imposed to pixel shaders.

Contributions: In this work we present a technique for solving Occlusion Culling in GPU, without the need for special hardware extensions or CPU read back. It includes a visibility test in the vertex shader of the application in order to discard those vertices that belong to occluded meshes. If the mesh is occluded then all its vertices can be discarded in the vertex shader, avoiding the rasterization step and the pixel computations. A previous step computes the visibility state of each mesh in the GPU and stores its result in an output texture called *Occlusion Map*. This result is acquired after performing a highly parallelized overlap and depth test comparison.

2 Related work

There is a great amount of research conducted on Occlusion Culling. A classification and overview of all these methods is presented by Cohen-Or et al. [6]. Among those techniques the ones that work in point-space are Hierarchical Z-Buffer (HZB) [7] and Hierarchical Occlusion Maps (HOM) [8].

On modern GPUs hardware occlusion queries [9] provide a built-in way to determine if a draw call contributes to the current frame, but suffer from latency and stalling effects due to the CPU read back. To address this issue temporal coherence techniques are applied [10], [11], but they require spatial hierarchies of objects to limit the number of issued queries.

Some newer hardware capabilities allow conditional rendering without CPU intervention like OpenGL conditional rendering which is implemented as GL NV conditional render [12] extension and DirectX 11 predicated rendering implemented as the ID3D11Predicate interface [13]. These methods determine whether geometry

should be processed or culled depending on the results of a previous draw call. Current hardware conditional rendering does not allow the GPU shaders to access the occlusion results, but Engelhard et al. [14] implement a method that allows this. Other authors [15], [16] also implement HZB on GPU using compute shaders.

More recently Nießner [17] proposes a patch primitive based approach to perform occlusion culling applying HZB and temporal coherence. In recent years, since CPUs increased the number of cores and the set of SIMD instructions were extended, some approaches perform point based Occlusion Culling such as HOM using highly optimized software rasterizers [18], [19], [20], [21].

3 Vertex Discard Occlusion Culling

3.1 Algorithm Overview

In our proposed method we perform a *from-point*, *image-precision* [6] occlusion culling process completely in GPU without the need for the CPU to read back the results. The method consists of a series of steps that must be followed by each frame to generate the *Occlusion Map*, perform the Visibility Test and obtain the Potentially Visible Set. Finally the method uses those results, already present in the GPU, to discard all the vertices of the occluded objects before they reach further stages of the pipeline. The steps are:

1. *Occludee Generation*: Select occluders and generate simplified volumes.
2. *Occlusion Map Generation*: Render occluder simplified volumes into the Occlusion Map Texture.
3. *Visibility Testing*: Determine which occludees are occluded and stored them in the Visibility Map.
4. *Vertex Discard*: Cull all the vertices that belong to invisible occludees.

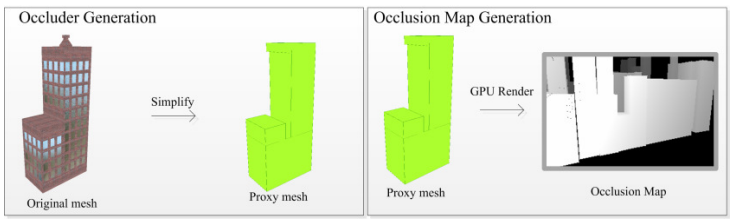


Fig. 2. The first step is to obtain the simplified occluders as proxy meshes. The second step is to render all proxy meshes to the Occlusion Map texture.

3.2 Occlusion Map Generation

The method begins Offline by creating a database of selected occluders that meet a predefined criteria [22], and storing the proxy meshes which are simplified, low-poly and conservative versions of the original occluders. These simplified occluders will be rendered faster than the original meshes, even if it behaves more conservatively. See Fig. 2.

In each frame, object-precision culling techniques such as Frustum Culling, PVS and Portal Culling [6] are applied to discard as many occluders as possible. With this obtained reduced subset of occluders we perform the first step of the method which is to render the proxy meshes into the *Occlusion Map*. This buffer stores the closest to camera depth values of every rasterized occluder and is implemented as a 32-bit floating point render target texture which is preferably a one fourth downscaled version of the screen framebuffer.

Unlike the HOM's Occlusion Map [8], our map does not contain opacity information, therefore the buffer is more similar to the HZB [7] which only stores the depth values of the occluders in each point, leaving the highest depth value to indicate no occluder presence.

The generation of the *Occlusion Map* is relatively inexpensive as the GPU massively parallel power is utilized to render the low-poly convex volumes of the proxy meshes and also because the pixel shader applied is extremely straightforward because it only outputs the depth value of each point.

3.3 Visibility Test

The core of this image based Occlusion Culling algorithm is to perform the Visibility Test for each selected occludee against the fusion of all the occluders represented by the *Occlusion Map*. Then it is used to determine whether the occludee geometry will continue along the pipeline or if it will be culled immediately. Visibility testing is performed by contrasting the points inside the occludee screen space bounding rectangle against the *Occlusion Map* depth values that contain the aggregated information of the occluders. In each frame, for every occludee in the viewing frustum, the algorithm performs a screen space projection of the occludee bounding box vertices. With those eight screen projected points, it determines the clipped 2D screen space bounding rectangle and finds the nearest from camera depth value of those extreme points. The resulting occludee bounding rectangle becomes a conservative superset of the actual pixels covered by the occludee (see Fig. 3). Afterwards, the visibility test determines if the occludee would actually contribute to the final image and starts by comparing all the depth values inside the occludee bounding rectangle against the ones in the *Occlusion Map*; when at least one point of the occludee is closer to the camera than the one stored in the same position in the *Occlusion Map*, the algorithm can now assume that the point is visible and therefore the whole occludee is considered potentially visible.

On the other hand, to determine that an occludee is completely culled, all the pixels must be examined exhaustively and proved to be farther than the values stored in the *Occlusion Map*.

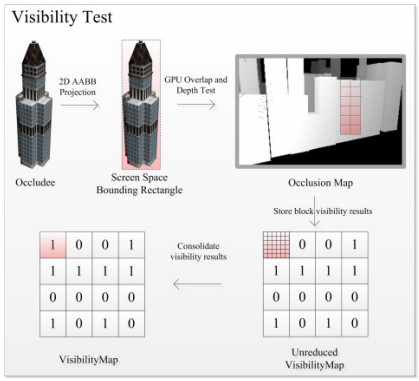


Fig. 3.The occludees in the scene are projected in 2D and the Bounding Rectangle is calculated. For each rectangle the algorithm performs the visibility test in GPU accessing the Occlusion Map, storing the visibility result in the Visibility Map.

Some methods implement this overlap and depth test in CPU [19], [20], [23], [21], and others use special GPU hardware capabilities such as hardware occlusion queries [9] or the more modern predicate/conditional rendering [13], [12]. Our method manually computes the visibility result pixel by pixel utilizing GPU pixel shaders.

However as explained before, to actually conclude that a occludee is culled, we have to exhaustively test all the pixels inside the occludee bounding rectangle, resulting in $N \times M$ texture fetches to the *Occlusion Map*. As the screen space regions covered by the occludees get larger, the number of texels to fetch and test can reach very large numbers.

To accelerate this, some methods build a pyramid of downsized versions of the *Occlusion Map* where each increasing level is half the size in each dimension of the previous one. There are two approaches to utilize the pyramid, one is like the method used in HOM [8] and HZB [7] which they begin at some level of the pyramid depending on the occludee bounding rectangle size and have to go to the finest level to assure that the occludee is completely culled by the occluders.

The other approach [15], [16] only restricts itself to a selected level of the pyramid, limiting the possible number of texture fetches to a given constant to avoid the worst case scenario where they have to move to levels with greater detail. After implementing this last variation we found that the level of conservativeness was higher than expected for medium to large screen space occludees.

In this work we found that using a single level *Occlusion Map* of a fourth of the original screen buffer was a good tradeoff between number of texture fetches and level of conservativeness. In the next section we discuss the methods used to leverage the GPU hardware to perform this visibility test.

3.4 Block Subdivision

Despite having a downsized version of the *Occlusion Map*, performing all the $N \times M$ texture fetches in a single pixel shader execution does not perform as expected, because of the serial nature of the algorithm. In the best cases this inner loop could take only a few cycles whereas in other cases the same execution could take hundreds of thousands of cycles before it is finished.

For this reason, in our method the visibility test is parallelized taking advantage of the parallel execution nature of the pixel shaders, splitting the total region covered by each occludee into a series of fixed size blocks where each one only performs a maximum of 8×8 texture lookups to the *Occlusion Map* (see Fig. 4). This way each occludee bounding rectangle is split up in blocks that concurrently perform the visibility test by executing pixel shaders that return only two possible output colors: 0 meaning the block itself is completely occluded or 1 if the block is potentially visible.

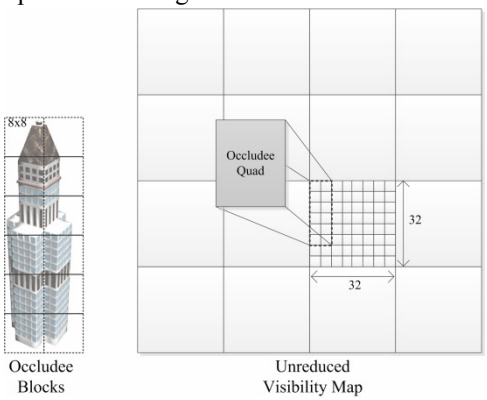
The output of each pixel shader goes to a rendering target texture called *Unreduced Visibility Map* (UVM) that holds the block visibility results one next to the other as seen in Fig. 5.

In order to simplify the way each region is assigned, every occludee is assumed to have a fixed number of blocks, regardless of its screen space size. In our study we determined that every occludee would have a preset number of 32×32 blocks assigned, resulting in a total of 1024 blocks. This gives us a maximum occludee screen size of 256×256 pixels and if the dimensions are larger than those, the occludee is simply considered potentially visible. To implement this algorithm using shader model 3 (without compute shaders), we carefully position a 32×32 pixel quad (GPGPU quad) and render it using a pixel shader that executes the visibility test code. Each pixel of this quad represents a block visibility test of the occluder. The shader gets the occludee bounding rectangle coordinates, depth value and the block number as parameters, and then executes the 8×8 pixels overlap and depth test.

```
Require: occludeeSize
Require: occludeePos {occludee AABB position}
Require: occludeeDepth
Require: occlusionMap
Require: pos {quad texture coordinates}
Require: quadSize
1:  base ← occludeePos × pos + quadSize × 8
2:  result ← 0 {not visible}
3:  for i = 0 to 8 do
4:          for j = 0 to 8 do
5:          p ← base + (i, j)
6:          depth ← read p from occlusionMap
7:              if occludeeDepth ≥ depth then
8:                result ← 1 {visible}
9:                break
10:             end if
11:         end for
12: end for
```

Fig. 4. Visibility test algorithm performed in a pixel shader

Using this block subdivision strategy, the visibility test is split into smaller task units and performed in parallel making use of the available GPU shader execution



cores. If all the blocks comprising the occludee rectangle output 0 values, then the whole occludee is considered culled, conversely when at least one of the blocks results visible the whole occludee is considered potentially visible.

Fig. 5. The occludee is split into 8x8 blocks, then each block performs the visibility test and stores the result into the Unreduced Visibility Map. Each occludee has a pre-assigned region of 32x32 blocks inside this Occlusion Map texture.

Nevertheless the visibility result of each occludee is not consolidated into a single value, but spread into a series of 32×32 matrices inside some region of the UVM. The next step of our method reduces each 32×32 occludee visibility result matrix into a consolidated *Visibility Map* that will hold the results of each visibility test one next to the other.

3.5 Visibility Map Reduction

In order to reduce the UVM and consolidate each 32×32 region into a single value, we need to determine if there is at least a non-zero value inside that matrix. To achieve this, we search for the maximum value of the matrix to see if there is any value other than zero. The search is done utilizing a parallel reduction approach with two rendering passes to limit the total number of operations. In the first pass we search the maximum value in each matrix column of 32 pixels and store it in an intermediate texture. In the second pass, we obtain the final *Visibility Map* looking for the maximum value in each row. Finally we end up with the *Visibility Map* containing the results of the occlusion culling process for each occludee tested in the current frame, which will be heavily accessed in the next step of our method.

3.6 Vertex Discard

This *Visibility Map* texture could be sent back to the CPU and processed there to avoid having to execute the draw calls to occluded objects; however this would produce a stalling effect on the GPU while sending the results back. To address this issue, we propose an asynchronous mechanism where the CPU does not need the results of the visibility test.

In our method the CPU always performs the draw calls for all the geometry that is potentially visible (the subset that passes frustum culling, portal culling, PVS, etc.), and the GPU is responsible for discarding the occluded geometry based on the *Visibility Map* content.

In our implementation we slightly modify the vertex shader that performs the *World-View-Projection* transformation as seen in Fig. 6. Before drawing an occludee, we send a parameter to the pixel shader indicating the *ID* of occludee that is about to be rendered. Based on that value, the vertex shader will perform a texture lookup in the *Visibility Map* to find the occlusion status for that particular occludee. If it is potentially visible, then the vertex shader does its usual computation letting the vertex continue throughout the pipeline. On the other hand, if the occludee is invisible we assign a negative z value to the output vertex so it can be culled by the GPU. This process is performed for every vertex that constitutes the occludee geometry.

Require: *vp* {Vertex 3D Position}
Require: *vMap* {Visibility Map}
Require: *i* {Occludee index}
1: *vis* ← read visibility info from *vMap* using *i*
2: **if** *vis* = 0 **then**
3: {Continue with normal vertex shader calculations}
4: **else**
5: *vp.z* = −1 {Discard vertex}
6: **end if**

Fig. 6. Vertex cull algorithm performed in a Vertex Shader.

4 Implementation and Results

Our method was implemented using C# 4.0 with DirectX 9 and Shader Model 3. We decided not to use newer shader models (with Compute Shader capabilities) so we could test in the current low-end commodity hardware. The implementation of our occlusion culling module was designed in a way that can be easily adapted to other graphics frameworks, where only certain parts have to be added or modified.

We tested our method in a densely occluded 3D city scene Fig. 7, composed of 210 meshes, adding up a total of 379,664 triangles. For this scene 258 occluder proxies were generated in Offline time based on the ideas presented by [22]. In order to analyze the algorithm performance, 15 representative scene View Points were taken, where in each position we compute the following occlusion metric:

$$value = \left(\frac{t - v}{t} \right) \times 100 \quad (1)$$

Where t is the total scene meshes and v is the total visible meshes. With this metric we can determine the percentage of meshes that were discarded by the GPU in each frame due to occlusion culling (see results in Fig. 8). These values are computed with Occlusion Culling deactivated and then with it activated. We also include the frames per second that resulted from rendering the scene using a pixel shader that alters the z value to produce a displacement mapping effect with Z-PrePass and with our Occlusion Culling method. On average our method increases the FPS around 20% compared to the Z-PrePass technique (see results in Fig. 8). The values were obtained using a PC with Intel Core i3 2.40GHz processor with 2GB RAM and Intel HD Graphics 3000 GPU.

5 Conclusions and Future Work

We have implemented a method that performs image space occlusion culling completely in GPU, taking advantage of its rendering power to build the *Occlusion Map* and leveraging its parallel architecture to perform the visibility test.

According to our results, this occlusion culling method is applicable in densely occluded scenes where pixel shaders are computationally expensive and specifically if they alter the default depth value of the fragments, like in [4] and [5]. Conversely we found that for scenes with lightweight pixel shaders and no depth overrides, our method performs similar to the GPU built-in Early-Z culling, making it suitable for mixed case scenarios.

As our implementation is based on Shader Model 3, it does not require special hardware requirements, beyond the vertex shader texture lookup capabilities present in most GPUs. However we found that in some older hardware, particularly those without Unified Shader architecture, the vertex texture lookup may downgrade the performance significantly [24]. It is also important to have some considerations before applying this technique. As all the occludees are sent to the GPU, no matter if they are occluded or not, there is a CPU-GPU bus bandwidth required to transfer the primitives to the graphics adapter. Moreover, as many other similar occlusion culling algorithms, the occluders have to be preprocessed in order to simplify the geometry into simpler conservative volumes.

Among the numerous enhancements to be made to our method, we would like to modify it to overcome the limitation of the 256×256 pixel size occludees and to explore built in hardware options to reduce the UVM, avoiding the current two rendering pass method.

Finally as newer versions of DirectX and OpenGL become available we could explore the option of implementing this method using compute shaders, orienting it to the work presented by Nießner[17] and Rákos[15]. We could also count the number of visible blocks in each occludee and utilize the results to determine some level of detail in geometry and pixel shaders.

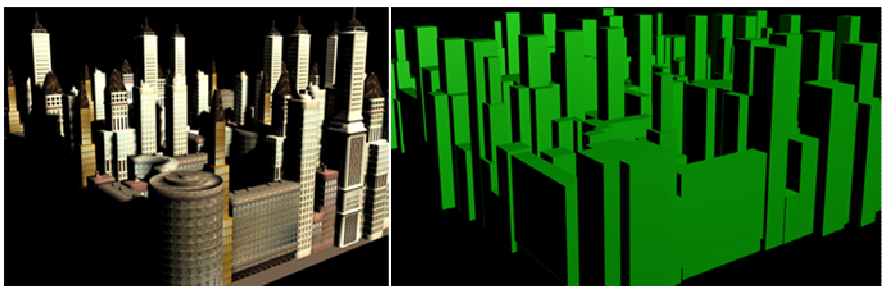


Fig. 7. Left: The 3D city scene used to test the algorithm. Right: The simplified occluder set used for Occlusion Culling

AKNOWLEDGMENTS

The authors would like to thank the GIGC Computer Graphics Research Group for supporting this research, and the Department of Information Systems Engineering for providing the support and funding. We also thank the Retrovia Project, specially Marta Garcen and Eva Ferrari (English node) for reviewing our work and to the Algebra and Analytic Geometry node for making this connection.

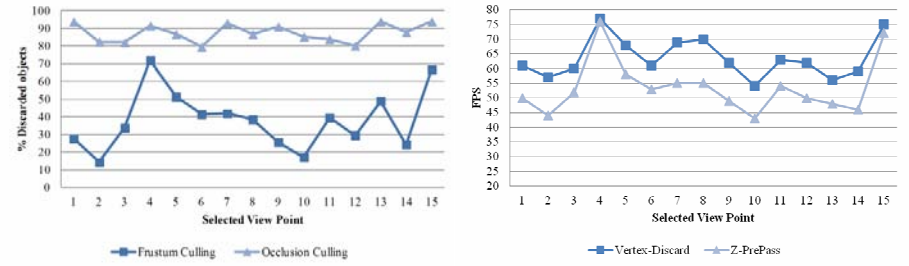


Fig. 8. Left: Discarded mesh percent, first with only Frustum Culling and then activating Occlusion Culling, at the fifteen different selected view points. Right: FPS rendering performance with Z-PrePass and then with Vertex-Discard Occlusion Culling activated, at the 15 different selected view points.

6 References

[1] Intel Corporation, "Early Z Rejection", http://software.intel.com/en-us/vcsource/samples/early-z-rejection, Accessed June. 2013.

[2] E. Haines and S. Worley, "Fast, low memory z-buffering when performing medium-quality rendering," J. Graph. Tools, vol. 1, no. 3, pp. 1–6, Feb. (1996).

[3] G. Riguer, "Performance optimization techniques for ati graphics hardware with directx 9.0," ATI Technologies Inc, (2002).

[4] V. Krishnamurthy and M. Levoy, "Fitting smooth surfaces to dense polygon meshes," in Proceedings of the 23rd annual conference on computer graphics and interactive techniques, ser. SIGGRAPH '96. New York, NY, USA: ACM , pp. 313–324. (1996)

[5] A. Lee, H. Moreton, and H. Hoppe, "Displaced subdivision surfaces," in Proceedings of the 27th annual conference on Computer graphics and interactive techniques, ser. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., pp. 85–94, (2000)

[6] D. Cohen-Or, Y. Chrysanthou, C. Silva, and F. Durand, "A survey of visibility for walkthrough applications," Visualization and Computer Graphics, IEEE Transactions on Visualization and Computer Graphics, vol. 9, pp. 412–431, (2003).

[7] N. Greene, M. Kass, and G. Miller, "Hierarchical z-buffer visibility," Anaheim, CA, pp. 231–238, (1993)

[8] H. Zhang, D. Manocha, T. Hudson, and K. Hoff, "Visibility culling using hierarchical occlusion maps." Los Angeles, CA: In Computer Graphics (Proceedings of SIGGRAPH 97), pp. 77–88, (1997)

[9] NVIDIA Corporation, "Nv occlusion query," http://www.opengl.org/ registry/specs/NV/occlusion query.txt, Accessed Mar. 2013.

[10] K. Hillesland, B. Salomon, A. Lastra, and D. Manocha, "Fast and simple occlusion culling using hardware-based depth queries," Techinical Report TR02-039, Dept. Comp. Sci., University of North Carolina, 2002.

[11] D. Staneker, D. Bartz, and M. Meissner, "Improving occlusion query efficiency with occupancy maps," in Proceedings of the 2003 IEEE Symposium on Parallel and Large-Data Visualization and Graphics, ser. PVG '03. Washington, DC, USA: IEEE Computer Society, p. 15, (2003)

[12] NVIDIA Corporation, "Nv conditional render," http://www.opengl.org/ , registry/specs/NV/conditional render.txt, Accessed Mar. 2013.

[13] Microsoft, "ID3D11Predicate interface," http://msdn.microsoft.com/en-us/library/windows/desktop/ff476577%28v=vs.85%29.aspx, Accessed Mar. 2013.

[14] T. Engelhardt and C. Dachsbacher, "Granular visibility queries on the gpu," Boston, pp. 161–167, (2009)

[15] D. R´akos, "Hierarchical-z map based occlusion culling," http://rastergrid. com/blog/2010/10/hierarchical-z-map-based-occlusion-culling/, Mar. 2013.

[16] N. Darnell, "Hierarchical z-buffer occlusion culling," http://www. nickdarnell.com/2010/06/hierarchical-z-buffer-occlusion-culling/, Accessed Mar 2013.

[17] M. Nießner and C. Loop, "Patch-based occlusion culling for hardware tessellation," in Computer Graphics International, (2012).

[18] W. Vale, "Practical occlusion culling in killzone 3," p. 49, (2011).

[19] J. Andersson, "Parallel graphics in frostbite-current & future," SIGGRAPH Course: Beyond Programmable Shading, (2009).

[20] Intel Corporation, "Software occlusion culling," http://software.intel. com/en-us/articles/software-occlusion-culling/, Accessed Jan. 2013.

[21] L. R. Barbagallo, M. N. Leone, M. M. Banquiero, D. Agromayor, and A. Bursztyn, "Techniques for an image based occlusion culling engine," in XVIII Argentine Congress on Computer Sciences, ser. CACIC 2012, Bahia Blanca, pp. 405–415, (2012)

[22] M. N. Leone, L. R. Barbagallo, M. Banquiero, D. Agromayor, and A. Bursztyn, "Implementing software occlusion culling for real-time applications," in XVIII Argentine Congress on Computer Sciences, ser. CACIC 2012, Bahia Blanca, pp. 416–426, (2012)

[23] H. Hey, R. F. Tobler, and W. Purgathofer, "Real-time occlusion culling with a lazy occlusion grid." London, UK, UK: Springer-Verlag, pp. 217–222, (2001)

[24] NVIDIA Corporation, "Geforce 8 and 9 series gpu programming guide," http://developer.download.nvidia.com/GPU Programming Guide/GPU Programming Guide G80.pdf, Accessed Mar. 2013.

A Multiple Object Tracking System Applied to Insect Behavior

Diego Marcovecchio[†‡], Natalia Stefanazzi[♯], Claudio Delrieux[†], Ana Maguitman[‡], and Adriana Ferrero[♯]

†Laboratorio de Ciencias de las Imágenes (LCI)
Departamento de Ingeniería Eléctrica y de Computadoras (DIEC)

‡Grupo de Investigación en Administración de Conocimiento y Recuperación de Información - LIDIA
Departamento de Ciencias e Ingeniería de la Computación (DCIC)

♯Laboratorio de Zoología de Invertebrados II
Departamento de Biología, Bioquímica y Farmacia (DBBF)

Universidad Nacional del Sur (UNS)
Av. Alem 1253, (B8000CBP), Bahía Blanca, Argentina
Tel: (0291) 459-5135 / Fax: (0291) 459-5136

Abstract. Segmentation and tracking of multiple objects is an extensively researched field among Image Sciences. Multiple object tracking is, in general, a very hard problem due to the great number of potencial issues that might arise (such as the sudden movement of the tracked objects, changes on the appearence of either the tracked objects or the background scene, bad-quality frames, occlusion between an object and the scene or between multiple objects, or camera movement). Normally, object tracking is performed on applications that require the object locations to perform calculations later. This paper describes the research, design and development of a system created in order to track multiple insects on recorded videos.

Keywords: Video Processing, Object Tracking, Computer Vision.

1 Introduction

Given the increasing ease of access to multimedia-recording devices (off-the-shelf and ready to use devices like tablets, smartphones and small movie cameras), it is now possible to record high-quality videos in an almost effortless manner; this motivates the development of software applications that automatically process such information.

Being aware of the complexity of the generic multiple object tracking problem, some context-dependent conditions are usually assumed in order to find practical but flexible solutions. In this paper, we will describe the methods used to create an application that performs the segmentation and tracking of

multiple objects (in particular, roaches running on Petri dishes) and will analyze several aspects on the path detected by each roach.

This work is the result of a joint-project between the *Laboratorio de Ciencias de las Imágenes* (IIIE-CONICET http://www.imaglabs.org), the *Grupo de Investigación en Administración de Conocimiento y Recuperación de Información* (http://ir.cs.uns.edu.ar), and the *Laboratorio de Zoología de Invertebrados II*, all of them from Universidad Nacional del Sur. The motivation for this work is that the first two groups develop an application to monitor the behavior of colonies of insects using the least invasive method possible, to test the effectiveness of several chemicals developed by the latter group. Another contribution and potential research direction is gaining insight for the development of new bioinspired algorithms based on ant-colonies techniques. The long-term and general goal is to develop an application that allows multiple detection and tracking of generic objects.

2 Related work

Multiple object tracking is a very complex task that requires the articulation of a pipeline with several sub-tasks to perform adequately. It is required to initialise proper regions of interest *(ROIs)*, to identify within them the desired targets, to perform a frame-by-frame following of the identified targets, to solve unexpected situations (like crossovers, superimpositions, and jerky movements) and to extract robust information regarding the individual trajectories of the targets.

Kim and Torralba proposed a system [10] that performs *ROI* detection very effectively. However, this system requires a set of different images to use as exemplar set, and in this work, we intend to perform *ROI* detection without training (i.e., we aim to detect our Region of Interest as soon as a video file is selected without any previous steps).

There are also some available programs that perform automatical or semi-assisted tagging of recognised actions in videos; for example, Takahashi's human action recognition in video surveillance systems [15], a traffic incident detection system [9], or an action-detection on tennis video recordings system [7]. However, these systems are not focused on tracking particular objects, and can only partially help to solve this problem.

Souded *et al.* presented an interesting feature-based particle filtering object-tracking system [13], but since it is focused on video-surveillance, it is designed to detect and track *every* moving object, while we are only interested in the insects inside a specific region of interest. Similarly, a more recent and robust system developed by Gao *et al.* [8] which also works with feature points based particle filtering was presented. However, the same problem remains: we are not interested in detecting *any* moving object, but only some insects inside a region. As we will see later, this approach can cause problems in our videos.

Agbinya and Rees used adaptive-color histogram-backprojection techniques to track multiple objects in surveillance and sports videos [1], but the system is not robust enough, and works with only short-lenght videos (around 10 seconds), while we need to track insects for approximately half an hour.

There is a known ant-tracking system created by Balch, Khan & Veloso[4] on the Carnegie Mellon University. Nevertheless, the application has several

limitations and problems such as unresolved occlusion between the ants and the recipient walls, losing the track whenever two of the ants get too close together, splitting of the bounding boxes due to specular reflexes of the ants, and losing the track whenever some ant stops moving and stays in place for a long time (because they consider them to blend into the background).

Finally, in a recently performed study [12], which required tracking of several dozen ants, each ant had a tiny label attached to its back. While the tracking system works correctly, we are trying to develop a less invasive method by trying to track the path of each insect without interacting physically with them.

By using more advanced video processing techniques, feature-detection and with some improvements on the heuristics, in addition to assuming some conditions that are not necessarily more restrictive, but do allow us to narrow the problem, we were able to eliminate or reduce drastically most of the mentioned limitations in Balch, Khan & Veloso's system.

3 The Application

We designed and developed an application that processes videos performing the detection, tracking and statistical analysis of insects (in this case, cockroaches) running on Petri dishes. The program was created using the free computer vision library `OpenCV`, and the `Qt` Framework. Next, we will describe each subsystem separately.

3.1 Videos

The videos used were recorded by the researchers at the *Laboratorio de Zoología de Invertebrados II* in order to evaluate the repelent action of essential oils extracted from a native northern plant from Argentina. Paper discs of 18 cm diameter were divided on two halves; one was treated with 1mL of essential oil, and the other remained non-treated. The paper discs were then placed inside of Petri dishes, covered with 10-cm plastic rings treated with vaseline in order to prevent the escape of the cockroaches. The videos were recorded in closed rooms, with controled moisture and temperature conditions, during 30 minutes.

3.2 ROI detection

In order to detect the Petri dish, which is our *Region of Interest*, first we get a Gaussian-filtered version of the frame using the following 5×5 convolution kernel:

$$k = \frac{1}{159}\begin{pmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{pmatrix}$$

Next, a Canny edge detection algorithm[6] is applied on the blurred frame: following a procedure analogous to Sobel, a pair of convolution masks are applied:

$$G_{\mathrm{x}} = \begin{pmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{pmatrix}$$

,

$$G_{\mathrm{y}} = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{pmatrix}$$

and the gradient strenght G and direction θ are found with:

$$G = \sqrt[2]{{G_{\mathrm{x}}}^2 + {G_{\mathrm{y}}}^2}$$

$$\theta = arctan\left(\frac{G_{\mathrm{y}}}{G_{\mathrm{x}}}\right)$$

rounding θ to 0, 45, 90 or 135. After applying *Non-Maximum Supression*, and with only candidate edges remaining, a final thresholding with hysteresis step is performed using default lower and upper threshold values (that can be modified by the program user if necessary).

Once a binary image containing the best candidate edges is obtained, contour detection is performed by using the alternative version of the border-following algorithm proposed by Suzuki and Abe [14] (which follows only the outermost borders of a binary image); after this step, only the dominant points of the curve are stored by applying the Teh-Chin chain approximation algorithm [16]. Finally, we decide which of the contours detected corresponds to the Petri dish by selecting the biggest contour that has a round enough shape. Note that for efficiency reasons, in case of camera motion only semi-automatical *ROI* reposition is applied (i.e., the *ROI* is not repositioned automatically frame-by-frame; the program user needs to click for it). Fig. 1 shows an example of the application determing the *ROI*.

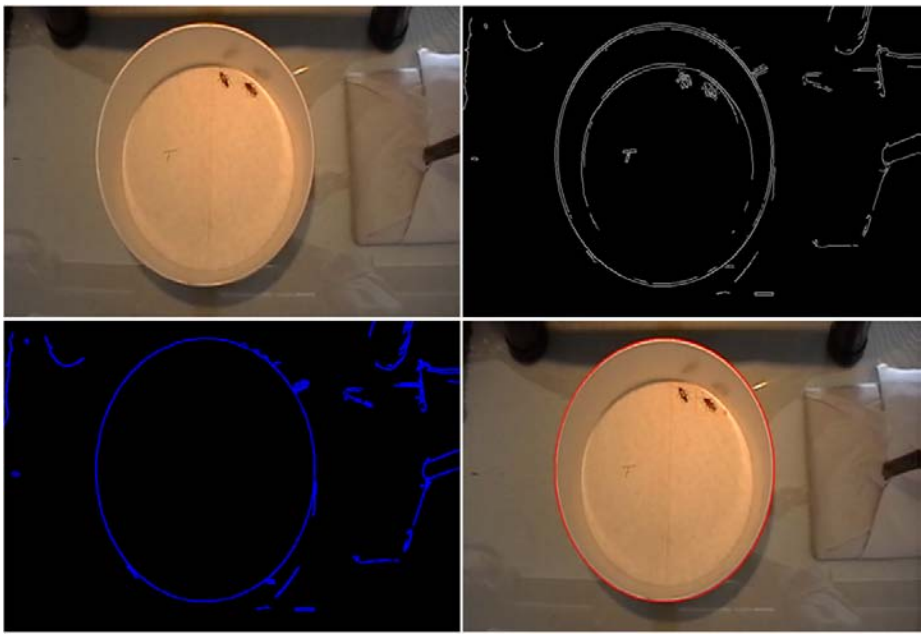
3.3 Segmentation

The application processes the video and initially recognizes the roaches by applying a *k-means* clustering algorithm [11] on the pixels inside the *ROI* that test positively in a comparison against a static color-characteristic centroid. Given the n positive pixels inside the *ROI*, k (number of roaches in the Petri dish) clusters are obtained by: first, randomly selecting k from the n pixels; second, associating each positive pixel with the closest of the selected k pixels, resulting in a Voronoi decomposition of the n pixels; third, the centroid of each of the k clusters becomes the new *mean*, and steps two and three are repeated until the variation epsilon $\varepsilon_{\mathrm{j_i}}$ in the iteration i of each centroid $\mathrm{P_j} = (\mathrm{X_j}, \mathrm{Y_j})$, with $j = 1..k$:

$$\varepsilon_{\mathrm{j_i}} = \sqrt{(X_{\mathrm{j_i}} - X_{\mathrm{j_{i-1}}})^2 + (Y_{\mathrm{j_i}} - Y_{\mathrm{j_{i-1}}})^2}$$

is small enough (in our case, $\varepsilon < 0.5$). As this is a heuristic algorithm, there is a chance it might not converge to the global optimum, depending on the

Figure 1 The ROI detection pipeline. From left to right and top to bottom: (a) the original video frame, (b) the Canny-Edge filtered image, (c) the detected contours on the Canny-Edge filtered image, and (d) the selected contour drawn on top of the original video frame.



initial clusters. Nevertheless, since the algorithm is fast enough (runs in polynomial smoothed complexity [2]), it is run multiple times with different starting conditions to check the correctness of the results.

Each insect (whose center is now defined by one of the k cluster centroids) is then trapped inside a bounding box. From now on, each of the roaches are identified frame-by-frame by using their bounding box, movement vector, and a path history that allows us to draw the trail of each insect.

Note that, due to this constraints, the system has two limitations at this moment: first of all, the number of insects needs to be known by the program user in order to apply the *k-means* algorithm. Second, since the color-characteristic centroid is initialized statically, the program would not be able to track white insects like *Myllocerus undatus Marshall* out-of-the-box. Nevertheless, it is important to notice that the system is flexible enough to allow easy user-settings to perform this operations in a semi-assisted way (i.e., to allow the user to select or specify the color-characteristic centroid, and then start to track the insects), and that with a few improvements all of this operations would be available automatically.

3.4 Tracking

In each frame, every bounding box surrounding an insect is analyzed pixel by pixel, applying for each one the comparison against the color-characteristic centroid once again. This way, the movement of the characteristic pixels inside each bounding box is detected. By applying erosion and dilation techniques in order to reduce video noise problems, each bounding box is adjusted according to the new mean position of the positive pixels found, and a new position is

added to the trail history.

The occlusion problems between insects mentioned on Balch, Khan and Veloso's system are partially solved by different methods; first of all, whenever two bounding boxes have overlapping pixels, this pixels are ignored. In the previously mentioned system, whenever two ants were too close together, the bounding boxes started capturing pixels from the other ants, and finally collided completely. In our approach, by discarding the overl apping pixels, each bounding box remains tracking only one insect. However, it could occur that not enough pixels are detected due to a large overlapping area between boxes; in this case, the previously described clustering algorithm is reapplied using every positive pixel inside the *ROI*. The algorithm, as mentioned, returns a new set of k points that correspond to the center of each of the roaches, but since several pixels might have been discarded in the previous frames due to being located in a region with overlapping bounding boxes, the new k centroids will probably not match the registered k bounding boxes centers perfectly, and will need to be adjusted to the new positions of the insects.

In order to decide which bounding box corresponds to each insect, a probabilistic model is used: the movement vector of each of the colliding roaches is obtained by analyzing the currently registered position of the roach, and its position 10 frames before. A new hypothetical position is obtained by calculating the mean position between the currently registered one, and the projected position of the roach (which is the currently registered position adding the corresponding movement vector). Finally, each of the bounding boxes are assigned to the free detected centroid that is closest to the new hypothetical position. Notice that this is another of the system's limitations: certain conditions (for example, two roaches staying in the same place, together, for a long enough amount of time) might cause a bounding box swap.

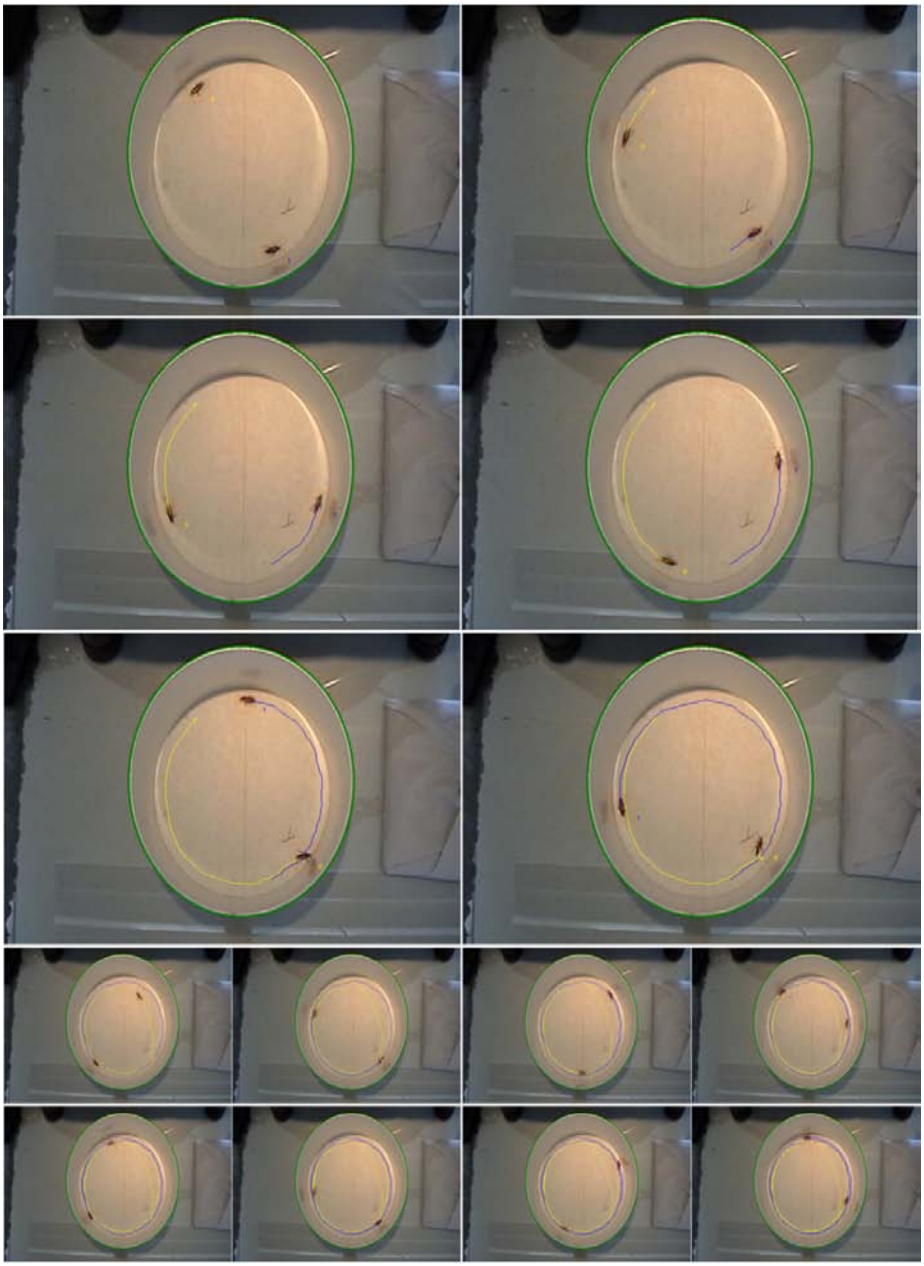
Another improvement in comparison to the ant-tracking system is that since we assume a constant amount of insects on the Petri dishes, and only check for positive pixels inside of the existing bounding boxes, these will never split into several ones. The ant-tracking system analyzed the difference between frames and placed a bounding box in each cluster of positive pixels, which caused new bounding boxes to appear in the specular reflexes on the Petri dish, and the system started tracking non-existent ants. The same problem occurs in the system developed by Gao *et al.*, because being a surveillance-oriented application, their system starts tracking every moving object. In our case, the bounding boxes simply keep following the actual insects.

Finally, the disappearing bounding boxes problem is solved, again, due to the fact that we adjust the position of the bounding boxes using the mean position of the characteristic pixels, instead of the difference between frames; in this way, the insects are not able to blend into the background. Fig. 2 shows the application tracking two insects for several frames.

3.5 Statistical analysis

Every piece of information gathered by the program is used to generate statistics that are later reported to the *Laboratorio de Zoología de Invertebrados II*; the most important is the time percentage spent on the treated and non-treated halves of the Petri dish, in order to determine if the developed insecticides are effective or not.

Figure 2 The application tracking two insects for several minutes. The trail left by each insect can be seen in a distinctive and unique color.



To accomplish this, each bounding box's center position is checked on every frame to see whether or not the insect has trespassed to the treated area of the Petri dish. The total number of trespassings can be compared to the total number of frames to obtain the time percentage spent on each half of the Petri dish. Notice that this is another of the system's limitations; tracking is not affected

by camera movement, and ROI reposition can be applied semi-automatically, but the statistical analysis will not be perfect if the camera is moved. However, this is not a hard-to-implement feature and could be added in future versions of the application.

In addition, because a history of the trail of each insect is stored, and it is possible to obtain a timestamp for each of them (or the time-delta between each of them), it is also plausible to analyze the tortuosity of the path of each insect by either the straightness, sinuosity, or fractal index [5].

4 Conclusions

The system currently detects and tracks effectively in every normal condition presented on the videos, being able to generate percentual statistics about the time spent by each roach on treated and non-treated regions of the Petri dish with great effectiveness. Once the target video is selected, the program works in a fully-automatic way, except for the semi-assisted *ROI* reposition (in case it is needed).

The application has also shown robustness when abrupt changes on the lightness ocurred on the room where the videos were recorded, and is in general a great improvement compared to the previously known ant-tracking system. In addition, due to the relative simplicity of the tracking algorithm, the application works fast enough to track insects in real time in 1280 x 720 videos at 30 frames per second, which most feature-based and complex tracking systems have serious trouble with. Nevertheless, the system presents some limitations. The color of the insects is defined statically, and whenever two insects occupy the same space during a large amount of time, the application may confuse them and could potentially swap the bounding boxes. Similarly, abrupt camera motion requires a user response in order to explicitely ask for a *ROI* repositioning, and makes the statistical analysis less effective.

Currently, the system is being used on a daily basis to test the effectiveness of the essential oils over dozens of videos.

5 Future work

There are several features we would like to add to the system. First of all, it would be desirable to perform the segmentation of each insect without using a statically defined characteristic color. It would also be useful to add sanity checks in order to test if the insects are effectively trapped inside their bounding boxes, and if the *ROI* is correctly positioned at some time. *ROI* tracking to detect camera motion is another possibility. Dynamic detection of the treated and non-treated areas of the Petri dish would eliminate the camera motion constraints. Adding feature-based techniques to the tracking system would make the application even more robust. Roaches collisions could be resolved in a more complex and robust way (for example, an implementation of the Minimum Cost Bipartite Matching algorithm [3]). And finally, a different clustering algorithm could be applied to check how many insects are present in the video, instead of using this knowledge beforehand to apply k-means.

References

[1] Johnson I Agbinya and David Rees. Multi-object tracking in video. *Real-Time Imaging*, 5(5):295 – 304, 1999.

[2] David Arthur, Bodo Manthey, and Heiko Röglin. k-means has polynomial smoothed complexity. *CoRR*, abs/0904.1113, 2009.

[3] Hari Asuri, Michael B. Dillencourt, David Eppstein, George S. Lueker, and Mariko Molodowitch. Fast optimal parallel algorithms for maximal matching in sparse graphs. Technical Report 92-01, Univ. of California, Irvine, Dept. of Information and Computer Science, Irvine, CA, 92697-3425, USA, 1992.

[4] Tucker Balch, Zia Khan, and Manuela Veloso. Automatically tracking and analyzing the behavior of live insect colonies. In *Proceedings of the fifth international conference on Autonomous agents*, AGENTS '01, pages 521–528, New York, NY, USA, 2001. ACM.

[5] Simon Benhamou. How to reliably estimate the tortuosity of an animal's path:: straightness, sinuosity, or fractal dimension? *Journal of Theoretical Biology*, 229(2):209 – 220, 2004.

[6] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986.

[7] Q. Huang G. Zhu, C. Xu. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *in Proc. ACM Multimedia, 2006*, pages 431–440, 2006.

[8] Tao Gao, Guo Li, Shiguo Lian, and Jun Zhang. Tracking video objects with feature points based particle filtering. *Multimedia Tools and Applications*, 58(1):1–21, 2012.

[9] Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, and Masao Sakauchi. Incident detection at intersections utilizing hidden markov model. In *6th World Congress on Intelligent Transport Systems*, pages 1–10, 1999.

[10] Gunhee Kim and Antonio Torralba. Unsupervised Detection of Regions of Interest using Iterative Link Analysis. In *Annual Conference on Neural Information Processing Systems (NIPS 2009)*, 2009.

[11] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[12] Danielle P. Mersch, Alessandro Crespi, and Laurent Keller. Tracking Individuals Shows Spatial Fidelity Is a Key Regulator of Ant Social Organization. *Science*, 340(6136):1090–1093, May 2013.

[13] Malik Souded, Laurent Giulieri, and Francois Bremond. An Object Tracking in Particle Filtering and Data Association Framework, Using SIFT Features. In *International Conference on Imaging for Crime Detection and Prevention (ICDP)*, London, Royaume-Uni, November 2011.

[14] Satoshi Suzuki and KeiichiA be. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32 – 46, 1985.

[15] M. Takahashi, M. Naemura, M. Fujii, and S. Satoh. Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 9–16, 2011.

[16] C. H. Teh and R. T. Chin. On the detection of dominant points on digital curves. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(8):859–872, August 1989.

Segmentación de Imágenes de Ultrasonido por medio de un algoritmo rápido de contornos activos

Ignacio Bisso[1], Juliana Gambini[2,3]

[1]Instituto de Ciencias, Universidad Nacional de General Sarmiento
Juan M. Gutierrez 1150 Los Polvorines, Pcia. de Buenos Aires, Argentina
[2] Dpto. de Ingeniería Informática-Instituto Tecnológico de Buenos Aires
Madero 399 (C1106ACD) - Buenos Aires - Argentina
[3] Dpto. de Ingeniería en Computación-Universidad Nacional de Tres de Febrero
Valentín Gómez 4828, Caseros, Pcia. de Buenos Aires, Argentina
ibisso@ungs.edu.ar
mgambini@itba.edu.ar

Resumen El estudio e interpretación de imágenes de ultrasonido es un desafío en el área de procesamiento de imágenes, debido al ruido que este tipo de imágenes posee. En este trabajo se propone la utilización de un método de segmentación basado en conjuntos de nivel pero que no resuelve ecuaciones diferenciales sino que ajusta el contorno del objeto de interés por medio del intercambio de elementos entre dos listas de pixels vecinos. Se propone utilizar la distribución Gaussiana para modelar los datos provenientes de la imagen y estimar los parámetros correspondientes en cada paso del algoritmo, actualizando la información de la región que se desea segmentar. Con esta propuesta logramos una mejora significativa en la precisión del ajuste del borde del objeto de interés, comparado con el método original.

Keywords: Segmentación de imágenes de ultrasonido, Contornos Activos, Conjuntos de Nivel

1. Introducción

Las imágenes de ultrasonido son muy utilizadas en diagnóstico médico porque permiten el examen clínico del cuerpo de una persona en forma no invasiva. Las imágenes de ultrasonido o ecografías, son capturadas con un sitema de ondas de sonido de alta frecuencia que producen imágenes del interior del cuerpo humano en tiempo real y por lo tanto permiten, no solamente tomar imágenes estáticas de un órgano sino también observar su funcionamiento y movimiento a lo largo de un tiempo. Existen imágenes de ultrasonido 2D, 3D y 4D en las cuales el método de adquisición transforma las ondas en imágenes 2D, 3D o imágenes 3D a lo largo del tiempo, respectivamente [4]. Además, en los últimos años se han utilizado con grán éxito para operaciones e intervenciones guiadas por imágenes y para terapias especiales. Por estas razones es de suma importancia desarrollar

métodos automáticos de segmentación e interpretación de este tipo de imágenes, que además sean eficientes, muy precisos y funcionen en tiempo real.

Sin embargo, estas imágenes tienen los inconvenientes de que contienen ruido speckle (el cual es muy difícil de eliminar) y que poseen bajo contraste, que dificulta la tarea [15].

Existen diversos enfoques para la segmentación de imágenes ecográficas, basados en métodos clásicos de detección de bordes y aplicados a un problema clínico concreto, como por ejemplo el estudio de la próstata [20,25], estudios relacionados con ginecología y obstetricia [28,13,10,1] o investicaciones sobre las arterias coronarias y enfermedades del corazón [27,7]. En los artículos [23,15,14] se discuten varias alternativas de segmentación de este tipo de imágenes.

Otro enfoque para la detección de bordes y objetos de interés en imágenes son los métodos basados en contornos activos o también llamados *snakes* desde que fueron presentados en [9]. Estos métodos se basan en curvas que evolucionan por medio de la minimización de una ecuación diferencial, hasta adaptarse al borde del objeto de interés. Fueron posteriormente mejorados en artículos que utilizan conjuntos de nivel [19,5,16] y operadores morfológicos [3], todos ellos basados en la resolución de ecuaciones diferenciales en derivadas parciales, lo cual posee un alto costo computacional y no puede utilizarse en aplicaciones que requieran tiempo real.

Existen muchos métodos en la literatura para segmentación de imágenes ecográficas basadas en contornos activos y conjuntos de nivel, en el artículo [12] se utilizan contornos activos basados en *snakes*, en el artículo [8] se proponen modelos de curvas deformables, también con minimización de energía aplicados a estudios de las venas. En [11] se utilizan conjuntos de nivel combinados con conocimientos a priori de la forma del objeto de interés y también el espacio-escala. En el artículo [24] los autores presentan un enfoque que combina difusión anisotrópica con contornos activos basados en curvas B-spline, en el cual la evolución de la curva se realiza utilizando un coeficiente de variación local y el error de la norma de Turkey. En los artículos [26,6] se presentan métodos de detección de objetos en imágenes de ultrasonido utilizando contornos activos basados en la resolución de ecuaciones diferenciales en derivadas parciales.

Debido al ruido speckle que este tipo de imágenes posee, es muy útil utilizar distribuciones estadísticas para modelar los datos, por ejemplo, en el artículo [18] se utilza la dstribución Rayleigh para modelar los datos, combinada con un método de detección de bordes basado en máxima verosimilitud.

Todos estos métodos resultan robustos y eficientes, sin embargo tienen serias limitaciones en aplicaciones que requieran detección de contornos en tiempo real y por lo tanto no pueden ser aplicados al diganóstico médico utilizando secuencias de imágenes.

Por otro lado, en los artículos [21,22], los autores proponen un método de seguimiento de objetos en video, que utiliza la teoría de conjuntos de nivel pero en el cual no es necesario resolver ecuaciones diferenciales en derivadas parciales, sino que la evolución se realiza por medio de intercambio de pixeles. Este método también puede utilizarse para segmentación de una imagen estática y es la

base del presente trabajo para segmentar las imágenes de ultrasonido. Una variación de este método se presenta en [2] donde los autores combinan la evolución por medio del intercambio de pixels con la representación de curvas B-Spline y realizan una aplicación a imágenes médicas de diferente tipo.

Las imágenes de ultrasonido o ecográficas, plantean importantes desafíos a los algoritmos de detección de contornos debido a que las mismas presentan bajo contraste y son muy ruidosas. Por esa razón, el algoritmo de [21] como fue planteado originalmente para seguimiento de objetos en video que no posean ruido, no sirve para ajustar el borde de un órgano u objeto imágenes ecográficas.

En este trabajo se propone mejorar el algoritmo de ajuste de contornos mediante intercambio de pixeles presentado en [21] agregándole la capacidad de considerar información de los pixeles fuera del objeto y de la región que se desea ajustar y modelando los datos con una distribución Gaussiana. Presentamos las dificultades que posee el método [21] cuando es aplicado a imágenes de ultrasonido y se proponen modificaciones para mejorar su comportamiento cuando es aplicado a este tipo de imágenes.

Este trabajo está compuesto de la siguiente manera: en la Sección 2 se presenta una síntesis del método original de detección de bordes que es utilizado como base en este trabajo. En la Sección 3 se explican los problemas que aparecen en su aplicación a imágenes de ultrasonido y las modificaciones realizadas al algoritmo original, constituyendo el aporte más importante de este trabajo. En la Sección 4 se muestran los resultados obtenidos al aplicar el algoritmo nuevo. Finalmente, en la Sección 5 se extraen conclusiones y se presentan trabajos futuros.

2. Intercambio de pixels para detección de bordes

Con el objetivo de que este trabajo sea autocontenido, explicamos en esta sección el método de detección de bordes utilizado, presentando los aspectos teóricos, la forma de representar las curvas y los pasos del método. Para mayor información sobre este tema ver [19,21].

2.1. Representación de las regiones

Sean $\{\Omega_1, \Omega_2, \ldots, \Omega_M\}$, con $\Omega_i \cap \Omega_j = \emptyset$ si $i \neq j$ el conjunto de M regiones de interés, las cuales son seleccionadas inicialmente por el usuario, el fondo $\{\Omega_1 \cup \Omega_2 \ldots, \Omega_M\}^C$ y el conjunto de contornos de las M regiones $\{C_1, C_2, \ldots, C_M\}$. Cada una de las regiones tiene asociado un vector $\Theta^i = (\theta_1^i, \theta_2^i, \ldots, \theta_n^i)$, con $i \in \{1, 2, \ldots, M\}$ que contiene información representativa de la región y por eso se dice que es su *vector de características*. Ejemplos de este vector son la tupla RGB correspondiente al color, o el vector cuyas componentes son los parámetros estimados de una distribución de probabilidad con la que se modela los datos provenientes de una región. Dado un pixel $\mathbf{x}$, la probabilidad de que el mismo pertenezca a la región i se denota como $P(\Theta^i(\mathbf{x}) \mid \Omega_i)$.

Dada una escena, las M curvas que corresponden a las regiones de interés están dadas por el conjunto $\{C_0, C_1, \ldots, C_M\}$ que minimiza la función

$$E(C_0, C_1, \ldots, C_M) = -\sum_{i=0}^{M} \int_{\Omega_i} \log(p(\Theta^i(\mathbf{x}) \mid \Omega_i))d\mathbf{x} + \lambda \int_{C_i} ds \tag{1}$$

El primer término es menor en la medida en que la probabilidad de que el pixel pertenezca a la región es alta. El segundo término es menor cuando las curvas tienen menos pixeles; esto hace que se vean favorecidas aquellas curvas más suaves.

2.2. Representación y Evolución de las Curvas

Con el fin de simplificar la explicación, se analiza el caso de una única curva C que rodea a la región Ω_1; el fondo es denotado con Ω_0.

Ecuaciones paramétricas La curva cerrada C que representa el borde del objeto de interés puede escribirse en forma paramétrica:

$$C(s) = (x(s), y(s)) \text{ con } 0 \leq s \leq S \text{ y } C(0) = C(S) \tag{2}$$

Dada una curva inicial, consideramos su evolución dentro de la imagen agregando una variable temporal a la ecuación de la curva, con lo que se tiene $C(s,t)$. Si la evolución de la curva se realiza por una fuerza G, se tiene:

$$\begin{array}{ll} \frac{\partial C}{\partial t}(s,t) = G(s,t).\hat{N} & s \in [0,1], t \in (0,\infty) \\ C(s,0) \ = C_0(s) & s \in [0,1] \end{array} \tag{3}$$

donde $\hat{N}$ es el vector unitario normal a la curva.

Conjuntos de nivel En este caso, la curva se expresa en forma implícita como el conjunto de nivel cero de una función $\varphi : \mathbb{R}^2 \to \mathbb{R}$. Esta función debe verificar:

$$\begin{array}{l} \varphi(\mathbf{x}) < 0 \ \text{ si } \ \mathbf{x} \in \Omega_1 \\ \varphi(\mathbf{x}) > 0 \ \text{ si } \ \mathbf{x} \in \Omega_0 \\ \varphi(\mathbf{x}) = 0 \ \text{ si } \ \mathbf{x} \in C \end{array} \tag{4}$$

Al tener en cuenta el paso del tiempo en la evolución de la curva se agrega un parámetro más a la expresión, obteniendo $\varphi(\mathbf{x}, t)$. Considerando que la evolución está gobernada por un campo $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, se obtiene la siguiente expresión:

$$\begin{array}{ll} \frac{\partial \varphi}{\partial t} + \mathbf{F} \cdot \nabla\varphi = & 0 \\ \varphi(\mathbf{x}, 0) & = \varphi_0 \end{array} \tag{5}$$

donde φ_0 es la curva inicial.

El campo $\mathbf{F}$ puede descomponerse en sus componentes tangencial y normal de la siguiente manera: $\mathbf{F} = \mathbf{F}_N\hat{N} + \mathbf{F}_T\hat{T}$ donde $\hat{N}$ y $\hat{T}$ son los vectores unitarios normal y tangencial a la curva, respectivamente. Es posible demostrar que la evolución en la componente tangencial es despreciable, con lo cual la Ec. 5 queda

$$\begin{array}{rl} \frac{\partial\varphi}{\partial t} + \mathbf{F}_N\hat{N}\cdot\nabla\varphi = & 0 \\ \varphi(\mathbf{x},0) & = \varphi_0 \end{array} \tag{6}$$

El vector $\hat{N}$ puede escribirse como $\hat{N} = \frac{\nabla\varphi}{|\nabla\varphi|}$ y por lo tanto

$$\hat{N}\cdot\nabla\varphi = \frac{\nabla\varphi}{|\nabla\varphi|}\cdot\nabla\varphi = \frac{|\nabla\varphi|^2}{|\nabla\varphi|} = |\nabla\varphi| \tag{7}$$

Así pues, se obtiene para la Ec. 6

$$\begin{array}{rl} \frac{\partial\varphi}{\partial t} + \mathbf{F}_N|\nabla\varphi| = & 0 \\ \varphi(\mathbf{x},0) & = \varphi_0 \end{array} \tag{8}$$

2.3. Algoritmo rápido de detección de bordes

En lo que sigue explicamos la evolución de la curva por intercambio de pixels, la cual está inspirada en la teoría explicada en las secciones anteriores; utilizando una sola región de interés en la escena para su mejor comprensión.

Dada una imagen I compuesta por una región de interés Ω_1, cuyo borde se quiere encontrar, y por la región Ω_0 correspondiente al fondo, donde $\Omega_0 \cup \Omega_1 = I$ y $\Omega_0 \cap \Omega_1 = \emptyset$. Cada una de las regiones está caracterizada por los parámetros Θ_m de una función de distribución $P(\mathbf{x}|\Theta_m)\,(m = 0, 1)$.

Se elige una función $\phi : R^2 \to R$ tal que

$$\begin{array}{c} \phi(\mathbf{x}) < 0 \; si \; \mathbf{x} \in \Omega_1 \\ \phi(\mathbf{x}) > 0 \; si \; \mathbf{x} \in \Omega_0 \end{array}$$

Sea C_1 el borde de la región Ω_1 que se quiere encontrar. Se definen dos listas de pixels vecinos L_{in} y L_{out} de la siguiente forma:

$$L_{in} = \{\mathbf{x}|\phi(\mathbf{x}) < 0 \;\text{ y }\; \exists\,\mathbf{y} \in N_4(\mathbf{x}) \;,\; \phi(\mathbf{y}) > 0\}$$

$$L_{out} = \{\mathbf{x}|\phi(\mathbf{x}) > 0 \;\text{ y }\; \exists\,\mathbf{y} \in N_4(\mathbf{x}) \;,\; \phi(\mathbf{y}) < 0\}$$

donde $N_4(\mathbf{x}) = \{\; \mathbf{y}/\,|\mathbf{x} - \mathbf{y}| = 1\}$ es el conjunto de pixels 4-vecinos de $\mathbf{x}$.

L_{in} y L_{out} son los bordes interno y externo del contorno C_1, respectivamente. La Figura 1 muestra un objeto en una imagen y sus bordes interno y externo, representados en diferentes tonos de gris.

En este método la ecuación de evolución está dada por

$$\phi_t(\mathbf{x}, t) = |\nabla\phi(\mathbf{x}, t)|\,(F_d + F_s) \tag{9}$$

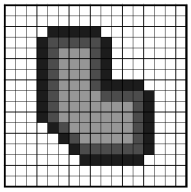


Figura 1. Objeto en una imagen, el borde interno (gris claro) y el externo (gris oscuro).

donde

$$F_d(\mathbf{x}) = \log\left(P\left(\theta^1(\mathbf{x})|\Omega_1\right)/P\left(\theta^0(\mathbf{x})|\Omega_0\right)\right) \tag{10}$$

es el término que hace que la curva se ajuste al borde del objeto de interés y

$$F_s(\mathbf{x}) = -2\lambda\kappa(\mathbf{x}) \tag{11}$$

siendo κ la curvatura, es el término que corresponde a realizar el suavizado.

Un pixel $\mathbf{x}$ es interior si $\mathbf{x} \in \Omega_1$ y $\mathbf{x} \notin L_{in}$. Un pixel $\mathbf{x}$ es exterior si $\mathbf{x} \in \Omega_0$ y $\mathbf{x} \notin L_{out}$. Entonces se define ϕ de la siguiente manera:

$$\phi(\mathbf{x}) = \begin{cases} 3 & si\ \mathbf{x}\ es\ un\ pixel\ exterior \\ 1 & si\ \mathbf{x} \in L_{out} \\ -1 & si\ \mathbf{x} \in L_{in} \\ -3 & si\ \mathbf{x}\ es\ un\ pixel\ interior \end{cases} \tag{12}$$

Luego se intercambian los pixels según un algoritmo de dos ciclos: en el primer ciclo se hace evolucionar el contorno siguiendo el signo de la fuerza $F_d(\mathbf{x})$. Como la fuerza $F_d(\mathbf{x})$ depende de datos de la imagen y de la caracterización del objeto, hace que la curva inicial evolucione adaptándose al contorno del objeto. El segundo ciclo es idéntico al primero, pero utilizando la fuerza $F_s(\mathbf{x})$, que como depende de la curvatura, suaviza el contorno.

El algoritmo comienza con la especificación de una curva inicial dada por el usuario. El método consiste en expandir y contraer el contorno por medio del intercambio de los pixels entre los conjuntos L_{in} y L_{out} siguiendo los siguientes pasos,

1. Para cada $\mathbf{x} \in L_{out}$, si $F_d(\mathbf{x}) > 0$ entonces, eliminar $\mathbf{x}$ de L_{out} y agregarlo a L_{in}. Luego, $\forall \mathbf{y} \in N_4(\mathbf{x})$, con $\phi(\mathbf{y}) = 3$, agregar $\mathbf{y}$ a L_{out} y poner $\phi(\mathbf{y}) = 1$.
2. Después de aplicar el paso 1 algunos de los pixels $\mathbf{x}$ en L_{in} pasaron a ser interiores y por o tanto deben ser eliminados de L_{in} y modificar $\phi(\mathbf{x}) = -3$.
3. Para cada $\mathbf{x} \in L_{in}$, si $F_d(\mathbf{x}) < 0$ entonces, eliminar $\mathbf{x}$ de L_{in} y agregarlo a L_{out}. Luego, $\forall \mathbf{y} \in N_4(\mathbf{x})$, con $\phi(\mathbf{y}) = -3$, agregar $\mathbf{y}$ a L_{in} y poner $\phi(\mathbf{y}) = -1$
4. Después de aplicar el paso 3 algunos pixels $\mathbf{x}$ se transformaron en pixels exteriores y por lo tanto deben ser eliminados de L_{out} y poner $\phi(\mathbf{x}) = 3$.

Notar que $F_d(\mathbf{x}) > 0$ implica que $P(\theta^1(\mathbf{x})|\Omega_1)/P(\theta^0(\mathbf{x}|\Omega_0) > 0$ y por lo tanto $\mathbf{x}$ es un pixel interior de Ω_1 y $F_d(\mathbf{x}) < 0$ implica que $\mathbf{x}$ es un pixel exterior de Ω_1.

En el primer ciclo se ejecutan los pasos un número N_a de veces, donde $0 < N_a < \text{máx}\,(columns,\ rows)$. Este parámetro determina el límite de expansión o contracción que la curva puede tener en un entorno de la curva inicial. En el segundo ciclo se ejecutan los pasos un número N_g de veces y se produce el suavizado de la curva por medio de una convolución con un filtro Gaussiano, el cual imita el comportamiento de la evolución por curvatura dado en la Ec. 11 (ver [17]) y por lo tanto la velocidad de evolución es $F_s\,(\mathbf{x}) = G \otimes \phi(\mathbf{x})$.

3. Problemas y Soluciones

Modelamos los datos provenientes de la imagen con una distribución Gaussiana y estimamos los parámetros μ_{Ω_1}, σ_{Ω_1}, μ_{Ω_0} y σ_{Ω_0} correspondientes a cada una de las regiones Ω_1 y Ω_0, respectivamente utilizando los estimadores

$$\mu_{\Omega_i} = \frac{\sum_{x \in \Omega_i} x}{k_i} \text{ y } \sigma_{\Omega_i} = \frac{\sum_{x \in \Omega_i} (x - \mu_{\Omega_i})^2}{k_i - 1} \tag{13}$$

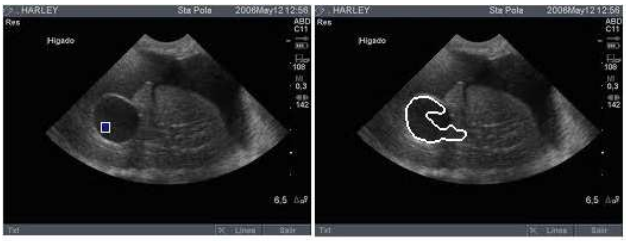
donde k_i es la cantidad total de pixels en la región Ω_i.

En el algoritmo original, se realiza la estimación de los parámetros tomando una muestra de la región de interés en forma supervisada por el usuario, eligiendo un rectángulo con el mouse, llamado R_1, dentro del objeto de interés y estimando los parámetros para el objeto utilizando el color de los pixels pertenecientes a R_1 y los parámetros del fondo utilizando el color del resto de los pixels en la imagen. La idea en ese caso es que todos los parámetros se estiman una sola vez. La Figura 4(a) muestra una imagen de ultrasonido correspondiente a un hígado donde se desea encontrar el objeto que aparece más oscuro. En la Figuras 2(a) y 2(b) pueden verse los resultados de aplicar el algoritmo original con diferentes regiones iniciales generando distintos resultados poco satisfactorios. En ambas Figuras, la región inicial se muestra del lado izquierdo. En esta figuras puede observarse que el resultado de aplicar el método depende en gran medida de la región inicial elegida. La Figura 3 muestra el resultado de aplicar el algoritmo a una imagen fetal de 7 semanas. Puede observarse que el resultado es mejor que los anteriores pero también tiene deficiencias en el ajuste.

La primer modificación que introducimos está relacionada con la estimación de los parámetros para la región del fondo. En el algoritmo original se estiman los parámetros μ_{Ω_0} y σ_{Ω_0} utilizando el complemento de la región R_1, lo cual considera todos los pixeles de la imagen, salvo los de R_1. Esto lleva a estimaciones erróneas en imágenes de ultrasonido debido a la naturaleza de las mismas, puesto que en este tipo de imágenes existen areas que no son parte de la imagen capturada mediante ultrasonido y que poseen pixeles de color negro como puede verse en la Figura 4(a). Incluir estos pixeles en la estimación no es conveniente porque que generan valores de los parámetros que no son adecuados para modelar la distribución de los pixeles del contorno de la región de interés.

En el presente trabajo se propone estimar los parámetros para el fondo μ_{Ω_0} y σ_{Ω_0} tomando una región en forma supevisada por el usuario, considerando solamente pixeles presentes fuera de la región de interés R_{Ω_0}. En la Figura 4(b)

(a) Región inicial R_1



(b) Región inicial R'_1

Figura 2. Resultado de aplicar algoritmo original utilizando dos regiones iniciales diferentes.

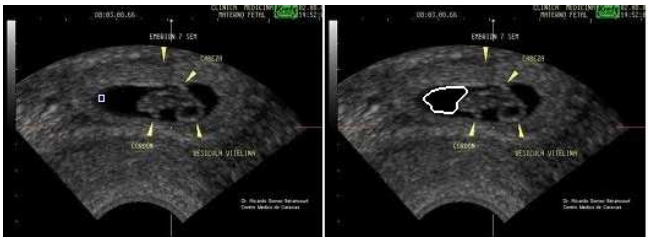


Figura 3. Resultado de aplicar el algoritmo original a una imagen fetal.

se muestra una región inicial delimitada por el usuario. Los valores de los estimadores obtenidos con esa imagen por el método original (MO) y el nuevo (MN) pueden verse en el Cuadro 1 donde se observa la diferencia entre ambos.

(a) Imagen Original (b) Area seleccionada

Figura 4. Imagen Original y área marcada en forma supervisada para estimar los parámetros de la distribución Gaussiana fuera del objeto de interés.

	μ_{Ω_0}	σ_{Ω_0}
MO	29	$35,8$
MN	82	$20,2$

Cuadro 1. Tabla de parametros estimados fuera de la región de interés utilizando dos regiones diferentes.

La segunda propuesta consiste en actualizar la estimación de los parámetros μ_{Ω_1} y σ_{Ω_1} en cada ciclo del algoritmo, es decir luego de cada intercambio de pixeles volver a calcular los parámetros con la nueva región encontrada. De esta manera a medida que el contorno evoluciona, también evolucionan los parámetros μ_{Ω_1} y σ_{Ω_1}. Si bien aumenta el costo computacional, realizar esta tarea incrementa la presición del resultado que se obtiene en el ajuste del contorno del objeto de interés, lo cual es de suma importancia en imágenes utilizadas en diagnóstico médico.

4. Resultados

En esta sección se muestran los resultados de aplicar algoritmo con las modificaciones propuestas. Las Figuras 5(a) y 5(b) corresponden al resultado de aplicar el algoritmo modificado utilizando dos regiones iniciales diferentes para estimar los parámetros del objeto de interés. Puede observarse que en ambos casos el resultado es un ajuste correcto al borde del objeto y que mejora notablemente los resultados obtenidos en las Figuras 2(a) y 2(b).

La Figura 6 muestra el resultado de aplicar el nuevo algoritmo a la imagen de la Figura 3, obteniendo un resultado más preciso.

(a) Región inicial R_1

(b) Región inicial R'_1

Figura 5. Resultado de aplicar el algoritmo con las modificaciones propuestas.

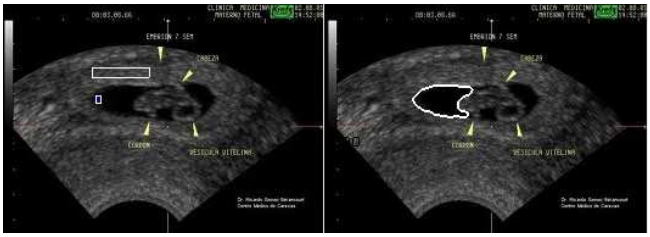


Figura 6. Resultado de aplicar el algoritmo con las modificaciones propuestas a la imagen de la Figura 3.

5. Conclusiones y Trabajos Futuros

En el presente trabajo se propone la utilización de un método de ajuste de contornos de objetos en imágenes de ultrasonido por medio de curvas. Como este tipo de imágenes son muy ruidosas y poseen bajo contraste, se propone utilizar la distribución Gaussiana para modelar los datos provenientes de la imagen, de forma tal que no solamente se tiene en cuenta la información del color del pixel sino también la forma en que los mismos están distribuidos. Esto provoca que se pueda encontrar el borde de los objetos dentro de la imagen con mayor precisión.

Como trabajos futuros pensamos utilizar otras distribuciones para modelar los datos con la distribución Γ o la $\mathcal{G}^0$, medir la bondad de la localización de la curva para poder comparar el algoritmo con otros métodos y aplicarlo a secuencias de imágenes evaluando su comportamiento en términos de tiempo.

Referencias

1. J. Anquez, E.D. Angelini, G. Grange, and I. Bloch. Automatic segmentation of antenatal 3-D ultrasound images. *IEEE Transactions on Biomedical Engineering*, 60(5):1388–1400, 2013.
2. O. Bernard and D. Friboulet. Fast medical image segmentation through an approximation of narrow-band B-spline level-set and multiresolution. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. ISBI '09.*, pages 45–48, 2009.
3. M. Bertalmío, G. Sapiro, and G. Randall. Morphing active contours: A geometric approach to topology-independent image segmentation and tracking. In *ICIP (3)*, pages 318–322, 1998.
4. S.L. Bridal, J.M. Correas, A. Saied, and P. Laugier. Milestones on the road to higher resolution, quantitative, and functional ultrasonic imaging. *Proceedings of the IEEE*, 91(10):1543–1561, 2003.
5. T. Chan and L. Vese. Active Contours without Edges. *IEEE Transactions on Image Processing*, 10(2):265–277, 2001.
6. M. Y. Choong, M. C. Seng, S.S. Yang, A. Kiring, and K. T K Teo. Foetus ultrasound medical image segmentation via variational level set algorithm. In *Third International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, pages 225–229, 2012.
7. L. Christodoulou, C.P. Loizou, C. Spyrou, T. Kasparis, and M. Pantziaris. Full-automated system for the segmentation of the common carotid artery in ultrasound images. In *5th International Symposium onCommunications Control and Signal Processing (ISCCSP)*, pages 1–6, 2012.
8. O. Husby and H. Rue. Estimating blood vessel areas in ultrasound images using a deformable template model. *Statistical Modelling*, 4:211–226, 2004.
9. M. Kass, A. Withkin, and D. Terzopoulos. Snakes: Active contour model. *International Journal of Computer Vision*, 1(1):321–333, March 1988.
10. Y. Li, Q. Huang, and L. Jin. A parameter-automatically-optimized graph-based segmentation method for breast tumors in ultrasound images. In *31st Chinese Symposium on Control Conference (CCC)*, pages 4006–4011, 2012.
11. N. Lin, W. Yu, and JS. Duncan. Combinative multi-scale level set framework for echocardiographic image segmentation. *Medical Image Analysis*, 7(4):529–537, 2003.

12. Ivana Mikic, Slawomir Krucinski, James D. Thomas, and Associate Member. Segmentation and tracking in echocardiographic sequences: Active contours guided by optical flow estimates. *IEEE Trans. Medical Imaging*, 17:274–284, 1998.
13. H. Neemuchwala, A. Hero, and P. Carson. Feature coincidence trees for registration of ultrasound breast images. In *International Conference on Image Processing*, volume 3, pages 10–13 vol.3, 2001.
14. J. A. Noble. Ultrasound image segmentation and tissue characterization. In *Proceedingsof the Institution ofMechanical Engineers, Part H: Journal of Engineering in Medicine*, 2010.
15. J.A. Noble and D. Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transaction on Meddical Imaging*, 25(8):987–1010, 2006.
16. S. Osher and N. Paragios. *Geometric Level Set Methods in Imaging, Vision and Graphics*. Springer, first edition, 2003.
17. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.
18. A. Sarti, C. Corsi, E. Mazzini, and C. Lamberti. Maximum likelihood segmentation of ultrasound images with Rayleigh distribution. *IEEE Transactions on Ultrasound Ferroelectric Frequency Control*, 52(6):947–960, 2005.
19. J.A. Sethian. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision and Materials Sciences*. Cambridge University Press, Cambridge, 1999.
20. F. Shao, K.V. Ling, W.S. Ng, and R.Y. Wu. Prostate boundary detection from ultrasonographic images. *Journal Ultrasound Medical*, 22:605–623, 2003.
21. Y. Shi and W.C. Karl. Real-time tracking using level sets. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 34–41, 2005.
22. Y.G. Shi and W.C. Karl. A real-time algorithm for the approximation of level-set-based curve evolution. *Image Processing*, 17(5):645–656, May 2008.
23. S. Sridevi and M. Sundaresan. Survey of image segmentation algorithms on ultrasound medical images. In *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013.
24. C. Tauber, H. Batatia, and A. Ayache. Robust B-spline snakes for ultrasound image segmentation. *Journal of Signal Processing Systems for Signal Image and Video Technology*, 54:159–169, 2009.
25. V. Wasson and B. Singh. Prostate boundary detection from ultrasound images using ant colony optimization. *International Journal of Research in Computer Science*, 1(1):39–48, 2011.
26. S. Yan, J. Yuan, and C. Hou. Segmentation of medical ultrasound images based on level set method with edge representing mask. In *3rd International Conference onAdvanced Computer Theory and Engineering (ICACTE)*, volume 2, pages 85–88, 2010.
27. X. Zhang, C.R. McKay, and M. Sonka. Tissue characterization in intravascular ultrasound images. *IEEE Transactions on Medical Imaging*, 17(6):889–899, 1998.
28. Y. Zimmer and S. Akselrod. Image segmentation in obstetrics and gynecology. *Ultrasound Medical Biolpgy*, 26(1):39–40, 2000.

Segmentación espectral de imágenes utilizando cámaras de tiempo de vuelo

Luciano Lorenti, Javier Giacomantone

Instituto de Investigación en Informática (III-LIDI),
Facultad de Informática - Universidad Nacional de La Plata - Argentina.
La Plata, Buenos Aires, Argentina.
{llorenti,jog}@lidi.info.unlp.edu.ar

Resumen. En este artículo se presenta un método de segmentación aplicable a imágenes adquiridas con cámaras de tiempo de vuelo (TOF). Las cámaras TOF generan dos imágenes simultaneas, una de intensidad y una de rango. El método propuesto modela ambas imágenes mediante una matriz de afinidad independiente para cada imagen. Transformando la imagen de rango y utilizando el criterio de cortes normalizados se optimiza la segmentación de los objetos de interés de la escena. El método mejora la segmentación en escenas donde la información de intensidad o de rango es insuficiente para obtener un separación adecuada de los objetos de interés. Se presentan resultados experimentales del método propuesto sobre imágenes simuladas e imágenes reales adquiridas por una cámara específica, y se derivan conclusiones a partir de los mismos.

Palabras clave: Segmentación, Imágenes de Rango, Cámaras de Tiempo de Vuelo, Agrupamiento Espectral

1 Introducción

El objetivo de un método de segmentación es dividir una imagen en sus partes constitutivas u objetos que la componen. La división depende del nivel de detalle requerido por el problema que se intenta resolver. Cuando una imagen de intensidad 2D brinda información limitada con respecto a la escena 3D que contiene los objetos a segmentar una alternativa posible es incorporar información de profundidad, la distancia de los distintos objetos que conforman la escena respecto a la cámara. En particular en este trabajo utilizamos una cámara de tiempo de vuelo,"Time of Flight" (TOF), que nos permite obtener imágenes de rango y de intensidad simultáneamente, la cámara utilizada es la MESA SR 4000 [1]. La SR 4000 es una cámara activa, utiliza su propia fuente de iluminación mediante una matriz de diodos emisores de luz infrarroja modulada en amplitud. Los sensores de la cámara detectan la luz reflejada en los objetos iluminados y la cámara genera dos imágenes. La imagen de intensidad es proporcional a la amplitud de la onda reflejada y la imagen de rango o distancia es generada a partir de la diferencia de fase entre la onda emitida y reflejada en cada elemento de la imagen

[2][3]. Las principales ventajas con respecto a otras técnicas de medición 3D es la posibilidad de obtener imágenes a velocidades compatibles con aplicaciones en tiempo real y la posibilidad de obtener nubes de puntos 3D desde un solo punto de vista[4][5]. Han sido utilizadas técnicas clásicas de segmentación directamente sobre imágenes de rango considerando distintas condiciones de ruido [6][7][8], [9] en distinto tipo de aplicaciones [10][11]. Recientemente han sido propuestas técnicas para segmentar objetos que operan sobre imágenes de rango e intensidad con el objetivo de definir bordes mas precisos en presencia, tanto de ruido como de oclusiones en distintos planos [12][13][14]. Han sido utilizados diversos algoritmos de agrupamiento tanto jerárquicos como particionales para abordar el problema de segmentar objetos en imágenes de intensidad. En particular el método propuesto, en primer término, asocia el problema de segmentación en el dominio de la imagen con la partición de un grafo. En segundo término optimiza el corte del grafo formulando la minimización necesaria para determinar el corte como un problema de autovalores generalizado [15][16][17]. Finalmente el método combina los resultados obtenidos del procesamiento de la imagen de intensidad y de la de rango mejorada [18], para obtener la segmentación final de los objetos de interés de la escena. La evaluación del método propuesto se realiza mediante dos tipos de datos, imágenes simuladas con todos los parámetros y factores de ruido controlados e imágenes reales adquiridas por la cámara SR 4000 en condiciones de operación real.
El artículo está organizado del siguiente modo, en la sección 2 se describe el método de partición mediante cortes normalizados y en la sección 3 se expone el método propuesto. En la sección 4 se presentan los resultados experimentales obtenidos. Finalmente en la sección 5 se presentan las conclusiones.

2 Cortes Normalizados en Segmentación de Imágenes

2.1 Particionado de Grafos

Un grafo G=(V,E) está formado por un conjunto de vértices V y un conjunto de aristas E que relacionan elementos de V. La teoría de grafos es usada, por lo general, en el modelado de problemas como tráfico de redes, circuitos eléctricos y redes de internet.
Con el objetivo de construir grafos a partir de imágenes, los vértices son generados a partir de los pixeles que la constituyen. Como cada uno de los elementos de la imagen contiene información de intensidad y posición, agrupar los pixeles de acuerdo a su semejanza y desemejanza puede lograr una correcta segmentación de la imagen. El conjunto de las aristas E está constituido por elementos que denotan la semejanza y desemejanza entre los pixeles. Como paso previo a la segmentación, es necesario construir un grafo pesado asignando a cada arista del conjunto E un peso $w(i, j)$, que resulta de evaluar la semejanza entre el pixel i y el pixel j. El valor de $w(i, j)$ aumenta con el grado de semejanza entre el pixel i y el pixel j. Denominamos corte de un grafo a la partición del mismo que se consigue removiendo las aristas de menor peso. Un arista con peso pequeño

indica un bajo grado de semejanza entre los pixeles que conecta. Por lo tanto para particionar el grafo en dos sub-grafos, el corte mínimo esta dado por:

$$cut(A,B) = \sum_{i \in A, j \in B} w(i,j)$$

donde A y B son dos subgrafos tales que $A \cup B = V, A \neq \emptyset, B \neq \emptyset$ y $A \cap B = \emptyset$. El criterio de corte mínimo, que fue introducido por Wu y Leahy[15], minimiza los posibles cortes máximos a través de los subgrafos. Para segmentar correctamente una imagen, se debe buscar el valor mínimo de corte entre todos los posibles subconjuntos del grafo, removiendo las aristas que involucren al corte mínimo y luego repetir el proceso de forma recursiva hasta que el valor de corte no supere cierto umbral.

2.2 Cortes Normalizados

La segmentación de imágenes utilizando el criterio de corte anterior otorga buenos resultados, pero favorece la aparición de sub-grafos de pocos pixeles aislados. Shi y Malik [16] propusieron un nuevo criterio, llamado cortes normalizados, que evita la aparición de estos conjuntos aislados. Definieron el criterio propuesto de la siguiente forma:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

donde $assoc(A,V) = \sum_{i \in A, j \in B} w(i,j)$ y $assoc(B,V)$ se define de manera análoga.

El criterio de cortes normalizados, a diferencia del criterio de corte mínimo, no considera la sumatoria de los pesos de las aristas que conectan los dos conjuntos sino que considera la proporción que esas aristas representan con respecto a la suma de los pesos de todas las aristas de los nodos del sub-grafo. Shi y Malik mostraron también que existe una relación entre el grado de asociación y disociación de los conjuntos, como se puede ver en la siguiente expresión:

$$\begin{aligned} Ncut(A,B) &= \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \\ &= \frac{assoc(A,V) - assoc(A,A)}{assoc(A,V)} + \frac{assoc(B,V) - assoc(B,B)}{assoc(B,V)} \\ &= 2 - \left(\frac{assoc(A,A)}{assoc(A,V)} + \frac{assoc(B,B)}{assoc(B,V)} \right) \\ &= 2 - Nassoc(A,B) \end{aligned}$$

Por lo tanto minimizar la disociación entre las particiones es equivalente a maximizar la asociación entre ellas.

2.3 Resolviendo los cortes normalizados como un problema de autovectores

Una de las ventajas más importantes para usar el criterio de cortes normalizados es que se puede obtener una buena aproximación de la particíon óptima de forma muy eficiente.
Sea $W_{ij} = w(v_i, v_j)$ la matriz de pesos del grafo y sea D la matriz diagonal de forma que $D_{ii} = \text{grado}(v_i) = \sum_{v_j \in V} w(v_i, v_j)$
Shi y Malik demostraron que una partición óptima se puede obtener calculando:

$$y = \arg\min Ncut = \arg\min_y \frac{y^T(D-W)y}{y^T D y}$$

donde y es un vector indicador binario que especifica a que grupo pertenece cada pixel.
Calcular el vector indicador es un problema NP-Completo para algunos tipos de grafos particulares, pero eliminando la restricción que y sea un vector indicador y resolviendo el problema en el dominio real, se puede hallar una aproximación a la solución discreta eficientemente. La ecuación anterior puede ser optimizada resolviendo el sistema de autovalores generalizado:

$$(D-W)y = \lambda D y$$

Shi y Malik demostraron que el segundo autovector de este sistema es la solución real del problema de cortes normalizados.

2.4 Ncut simultáneo de k-vías

Shi y Malik definieron el corte simultáneo de k-vías que da como resultado k segmentos en una sola iteración de la siguiente forma:

$$Ncut(A_1, A_2, \ldots, A_k) = \frac{cut(A_1,A_1)}{assoc(A_1,V)} + \frac{cut(A_2,A_2)}{assoc(A_2,V)} + \ldots + \frac{cut(A_n,A_n)}{assoc(A_n,V)}$$

Dado un vector indicador v de un sub-grafo A_j tal que

$$v_i = \begin{cases} \frac{1}{\sqrt{assoc(A_n,V)}} & \text{si } i \in A_j \\ 0 & \text{si } i \notin A_j \end{cases}$$

resulta que

$$v_i^T L v_i = \frac{cut(A_j,A_j)}{assoc(A_j,V)}$$

Sea H la matriz formada por k vectores indicadores puestos en columnas, minimizar $Ncut(A_1, A_2, \ldots, A_k)$ es equivalente a minimizar:

$$\min_{A_1, A_2, \ldots A_k} \mathrm{Tr}(H^T L H) \text{ sujeto a } H^T D H = I$$

Relajando el carácter discreto de la restricción y substituyendo $P = D^{\frac{1}{2}} H$ se obtiene el siguiente problema

$$\min_{P \in R^{x \times k}} \mathrm{Tr}(P^T D^{-1/2} L D^{-1/2} P) \text{ sujeto a } P^T P = I$$

Este problema de minimización de traza estándar es resuelto por la matriz P cuando esta contiene los primeros k autovectores, puestos en columnas, correspondientes a los autovalores mas pequeños de la matriz $LN = D^{-1/2} L D^{-1/2}$, llamada matriz laplaciana normalizada [19]. Resustituyendo $H = D^{-1/2} P$, se puede ver que estos autovectores son los autovectores generalizados correspondientes a los autovalores mas pequeños de $(D - W)u = \lambda D u$.

3 Método Propuesto

Sea $I(i)$ una imagen de intensidad y $R(i)$ una imagen de rango, ambas de dimensión $n \times m$.

1. A partir de la imagen $R(i)$ se genera una imagen de rango mejorada $NERI(i)$[18]

2. (a) Se construye la matriz de afinidad W_I de forma que

$$W_I(i,j) = e^{\frac{-\|F(i)-F(j)\|_2}{\alpha_I}} * \begin{cases} e^{\frac{-\|X(i)-X(j)\|_2}{\alpha_X}} & \text{si } \|X(i) - X(j)\|_2 < r \\ 0 & \text{c. c.} \end{cases}$$

   donde $X(i)$ es la locación espacial del nodo i y $F(i) = I(i)$.

   (b) Se construye la matriz de afinidad W_R de forma que

$$W_R(i,j) = e^{\frac{-\|F(i)-F(j)\|_2}{\alpha_I}} * \begin{cases} e^{\frac{-\|X(i)-X(j)\|_2}{\alpha_X}} & \text{si } \|X(i) - X(j)\|_2 < r \\ 0 & \text{c. c.} \end{cases}$$

   donde $X(i)$ es la locación espacial del nodo i y $F(i) = NERI(i)$.

3. (a) Se calcula la matriz laplaciana normalizada asociada a W_I

$$L_{NI} = D_I^{-1/2}(D_I - W_I)D_I^{-1/2}$$

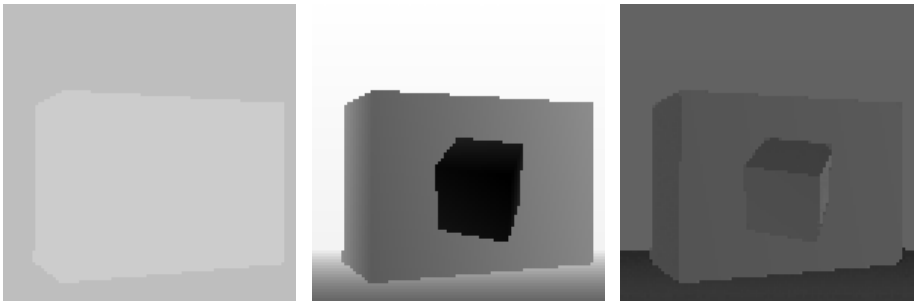
   (b) Se calcula la matriz laplaciana normalizada asociada a W_R

$$L_{NR} = D_R^{-1/2}(D_R - W_R)D_R^{-1/2}$$

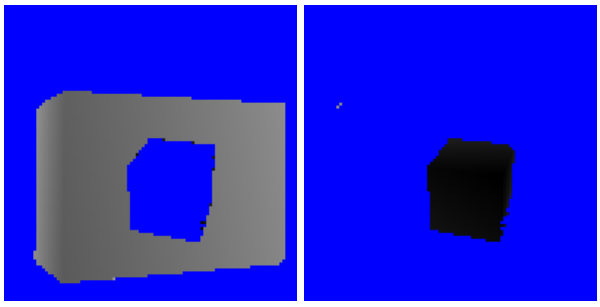
4. Se genera la matriz $H_I \in R^{k \times n}$ que contiene como columnas primeros k autovectores del sistema $L_{NI}u = \lambda u$. Se genera la matriz $H_R \in R^{l \times n}$ que contiene como columnas los primeros l autovectores del sistema $L_{NR}u = \lambda u$.
5. Se obtienen los autovectores correspondientes a las matrices laplacianas no normalizadas. $PI = D_I^{-1/2} H_I$ y $PR = D_R^{-1/2} H_R$
6. Se forma la matriz P_{IR} concatenando las matrices PI y PR
7. Para $i = 1, \ldots, n$, sea $y_i \in R^m$ los vectores correspondientes a la i-ésima fila de P_{IR}, se segmentan los puntos $y_i \in R^m$ con algoritmo k-medias en clusters $C_1, \ldots, C_k$.

4 Resultados Experimentales

Se presentan resultados experimentales del método propuesto aplicado a imágenes artificiales y a imágenes reales. Las capturas reales fueron obtenidas utilizando la cámara de tiempo de vuelo MESA SwissRanger SR4000 [1] y las capturas reales fueron simuladas utilizando el software Blensor [20].
En la figura 1 se puede visualizar el resultado de aplicar el algoritmo propuesto sobre una captura artificial.



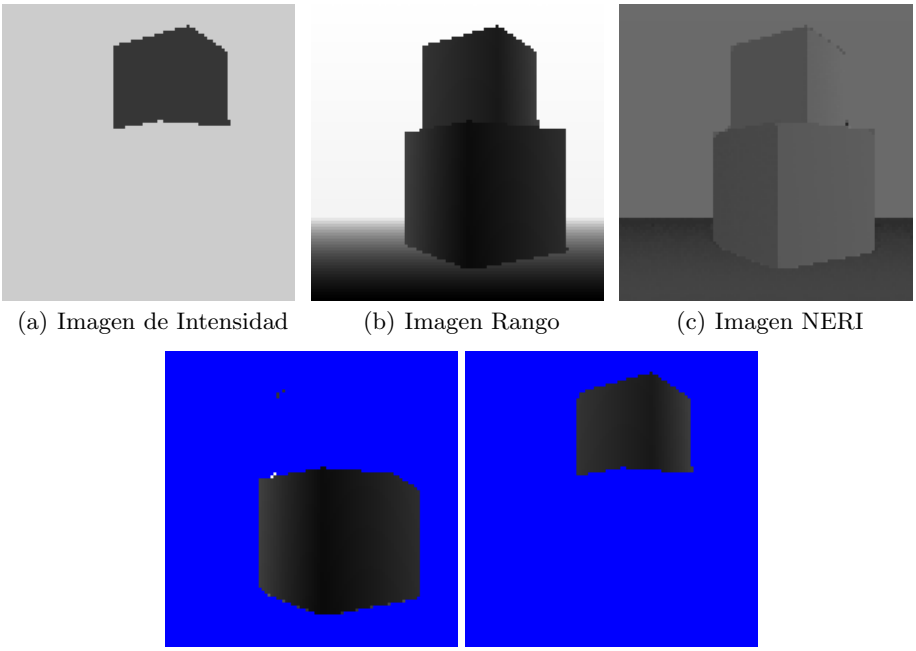
(a) Imagen de Intensidad (b) Imagen de Rango (c) Imagen NERI



(d) Resultado de la Segmentación

Fig. 1. Segmentación utilizando el método propuesto

La imagen 1(a) muestra dos objetos de niveles de intensidad iguales y ambos con niveles de intensidad próximos al del fondo. En la imagen 1(b) se puede observar que la información de distancia es útil para realizar una segmentación correcta. La imagen 1(c) muestra como la imagen de rango mejorada, incorpora información sobre las orientaciones de los objetos, conservando la información de distancia. Los parámetros utilizados para generar las matrices de afinidad y la cantidad de autovectores utilizados para la segmentación están detallados en la Tabla 1. Aplicando k-medias sobre el espacio generado por los autovectores, el algoritmo permite extraer correctamente las partes constitutivas de la imagen como se puede observar en las imágenes 1(d).



(a) Imagen de Intensidad (b) Imagen Rango (c) Imagen NERI

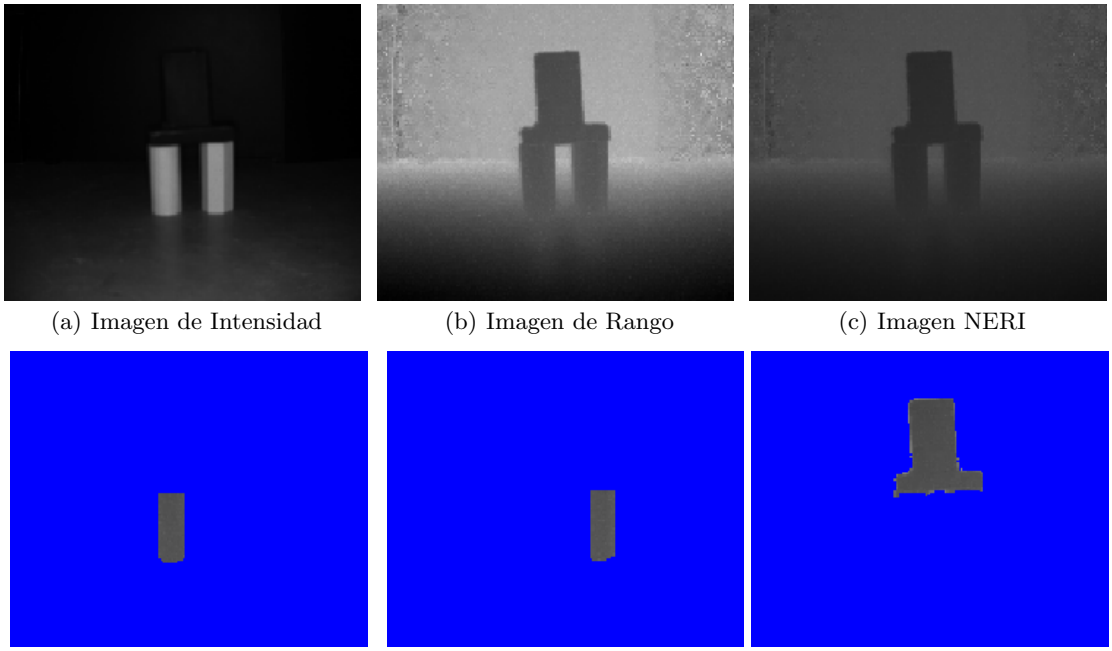
(d) Resultado de la Segmentación

Fig. 2. Segmentación utilizando el método propuesto.

En la figura 2 se puede visualizar el resultado de aplicar el algoritmo propuesto sobre otra captura artificial. La imagen 2(a) muestra como en la imagen de intensidad el segundo objeto tiene niveles de intensidades muy próximos a los del fondo y en la imagen 2(b) como los dos objetos que integran la imagen son difíciles de separar uno del otro. Al utilizar los parámetros especificados en la Tabla 1, el algoritmo puede combinar correctamente la información de las dos imágenes. Utilizando correctamente los datos de distancia para separar el fondo del objeto situado en la parte inferior de la imagen, y los datos de intensidad

para separar un objeto del otro como presenta la figura 2(d).

La figura 3 muestra el resultado de aplicar el algoritmo propuesto sobre una captura real. La imagen de amplitud 3(a) presenta 3 objetos sobre un fondo negro, todos a la misma distancia. Uno de los objetos tiene un nivel de intensidad similar al del fondo, lo que dificulta su segmentación. En la imagen de rango 3(b) los objetos se distinguen claramente del fondo pero no uno del otro. El método combina correctamente la información de ambas imágenes ruidosas para segmentar los tres objetos presentes en la escena, como se muestra en la figura 3.



(a) Imagen de Intensidad (b) Imagen de Rango (c) Imagen NERI

(d) Resultado de la Segmentación

Fig. 3. Segmentación utilizando el método propuesto sobre una captura real

	Imagen	Autovectores	r	α_I	α_X
Prueba 1	Intensidad	3	3	0.5	4
	Rango	3	3	2	4
Prueba 2	Intensidad	2	4	0.2	3
	Rango	2	4	12	2
Prueba 3	Intensidad	2	4	2	4
	Rango	3	4	190	2

Tabla 1. Parámetros

5 Conclusiones

En este artículo presentamos un método de agrupamiento aplicado a la segmentación de imágenes de rango. La descripción formal del método se plantea utilizando herramientas conocidas de álgebra lineal, simplificando así su implementación. Los resultados obtenidos tanto sobre imágenes de intensidad y rango simuladas como reales presentan resultados preliminares satisfactorios. El algoritmo combina adecuadamente la información provista por ambas imágenes incluso en presencia de ruido. Permite segmentar objetos con niveles de intensidad próximos, ubicados a distintas distancias, u objetos cercanos con niveles de intensidad diferentes o con orientaciones diferentes. Como trabajo futuro se propone un análisis detallado de la influencia de los parámetros de la función de pesos $w(i, j)$ en los autovectores de las matrices laplacianas. Otro aspecto para un trabajo de investigación futuro es modificar el método propuesto combinando la información de intensidad y profundidad en la función de pesos $w(i, j)$.

Referencias

1. M. Cazorla, D. Viejo, C.Pomares. Study of the SR 4000 camera. XI Workshop de Agentes Físicos, Valencia, 2010.
2. N. Blanc, T. Oggier, G. Gruener, J. Weingarten, A. Codourey, P. Seitz. Miniaturized smart cameras for 3D imaging in real time. IEEE Sensors, Vienna, Austria, 471-474, 2004.
3. J. Mure-Dubois, H. Hugli. Real-Time Sattering compensation for Time of Fligth cameras. International Conference of Vision Systems, 117-122, 2007.

4. A. A. Dorrington, C. D. Kelly, S. H. McClure, A. D. Payne, M. J. Cree. Advantages of 3D Time of Flight Range Imaging Cameras in Machine Vision Applications. 16^{th} Electronics New Zealand. Dunedin, New Zealand, 95-99, 2009.
5. F. Chiabrando, D. Piatti, F. Rinaudo. SR-4000 TOF Camera: Further Experimental Tests and First Applications to Metric Surveys. V Symposium on Remote Sensing and Spatial Information Sciences, Newcastle, UK, 38(5):149-154, 2010.
6. D. Ziou, S. Tabbone. Edge Detection Techniques - An Overview. International Journal of Pattern Recognition and Image Analysis, 8:537-559, 1998.
7. N. Otsu. A Threshold Selection Method from Gray.Level Histograms. IEEE Transactions on Systems, Man and Cynernetics, 9 (1):61-66, 1979.
8. D. Marr, E. Hildreth. Theory of Edge Detection. Proceedings of the Royal Society of London. Series B, Biological Sciences, 207 (1167):187-217, 1980.
9. J. Canny. A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8 (6):679-698, 1986.
10. R. Tanner, M. Studer, A. Zanoli, A. Hartmann. People detection and tracking with tof sensor. IEEE Conference on Advanced Video and Signal Based Surveillance, 356-361, 2008.
11. S. Caraian, N. Kirchner. Robust Manipulalability-Centric Object Detection in Time-of-Fligth Camera Point Clouds. Australian Conference on Robotics and Automation, 1-9, Brisbane, Australia, 2010.
12. R. Benlamri. Range image segmentation of scenes with occluded curve objects. Pattern Recognition Letters, 21: 1051-1060, 2000.
13. S. Oprisescu, C. Burlacu, A. Sultana. A new contour extraction algorithm for ToF images. 10^{th} International Symposium on Signals, Circuits and Systems (ISSCS), Bucharest, Romania, 1-4, 2011.
14. G. Danciu, M. Ivannovici, V. Buzuloiu. Improved Contours for ToF Cameras based on Vicinity Logic Operations. 12^{th} International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), 989-992, 2010.
15. Z. Wu, R. Leahy, An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.15, no.11, 10 1101-1113, 1993.
16. J. Shi, J. Malik, Normalized cuts and image segmentation.Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 731 - 737, 1997.
17. J. Shi, J. Malik, Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Inteligence, Vol. 22, No. 8, 888-905, 2000.
18. K. Pulli; M. Pietikainen. "Range image segmentation based on decomposition of surface normals". Proceedings of 8th Scandinavian Conference on Image Analysis. pp. 893-899, 1993.
19. R.K. Chung. Spectral Graph Theory. Am. Math. Soc, 1997.
20. M. Gschwandtner; R. Kwitt; A. Uhl, BlenSor: Blender Sensor Simulation Toolbox. Proceedings of 7th International Symposium In Advances in Visual Computing. ISVC 2011, 2011.

Procesamiento de Imágenes Muestrales de Fibra Textil de Origen Animal

Marcelo Arcidiácono[1] , Leticia Constable[1] , Juan Carlos Vázquez[1]

[1]Departamento de Ingeniería en Sistemas de Información, Facultad Regional Córdoba, Universidad Tecnológica Nacional, Maestro López esq. Cruz Roja Argentina, Córdoba.
{marceloarcidiacono,leticiaconstable,jcjvazquez}@gmail.com

Resumen. En el marco de la sustentabilidad productiva de fibra textil de origen animal, contar con un método ágil y seguro que permita obtener una medida de la calidad de la fibra, resulta sumamente valioso para los productores rurales. Una medida de calidad de la fibra textil puede obtenerse a partir del diámetro de la misma. El laboratorio del SUPPRAD lleva a cabo un procedimiento científico innovador en la obtención de este valor en la que se recoge un corte transversal del mechón en un portaobjetos, para ser fotografiado con equipo adosado a un microscopio. La imagen obtenida es manualmente procesada para determinar el diámetro promedio de fibra. Este procedimiento manual de medición, resulta lento, engorroso e introduce error por intervención humana. Se automatiza el procedimiento por medio de un software de procesamiento de imágenes. Se comentan los resultados obtenidos y se presentan las previsiones para la continuación de este trabajo.

Palabras Clave: Fibra textil, Calidad, Imágenes, Procesamiento Automático.

1 Introducción

La identificación y caracterización de pelos y fibras de origen animal (incluido el humano) o vegetal, adquiere una importancia relevante, por ejemplo, en la tipificación de la dieta en humanos y animales depredadores, en la confección de inventarios faunísticos [1], en la clasificación y en la estimación de abundancia de especies [2], en criminología [3], en la industria peletera [4] y por supuesto en el análisis y estudio de las fibras con propiedades y usos textiles industriales [5] e inclusive artesanales.

Más de un millón de pequeños productores de los Andes centrales de Sudamérica tienen alpacas y llamas como principal medio de subsistencia. Los animales proveen carne, leche, fibra, energía de transporte y guano y, además, constituyen un elemento importante de la identidad cultural de sus pueblos. Poblaciones específicas de estos camélidos califican para ser capturadas, esquiladas y liberadas generando un ingreso adicional a las comunidades en que viven. El aumento de la producción de fibras y demás productos de camélidos sudamericanos, a la vez de preservar un recurso genético animal crítico y los valores culturales asociados y mejorar la calidad de vida de muchos pequeños productores, debe ser parte de una estrategia global de inversión sostenida en investigación y desarrollo apropiados.

El Programa SUPPRAD de la Facultad de Ciencias Agropecuarias de la Universidad Católica de Córdoba, vinculado con instituciones y cooperativas agrícolas y ganaderas, tanto nacionales como internacionales [6], lleva adelante proyectos para identificar objetivos de mejoramiento de las cualidades de

sustentabilidad para la producción de pequeños rumiantes y camélidos en áreas desfavorecidas.

El proyecto de investigación RNA-SU del Departamento de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba, colabora con SUPPRAD en su objetivo de acercar la tecnología a productores de áreas desfavorecidas, desarrollando un software para medir la calidad de la fibra textil de origen animal. Esta herramienta se confecciona, sin descuidar la situación económico-cultural del usuario y atendiendo las exigencias de exactitud y precisión.

2 Ámbito del problema

En la República Argentina, el Programa Nacional "Fibras Animales" considera de gran valor la producción, comercialización e industrialización de lana, mohair, cashmere, llama, guanaco y vicuña [7].

La lana es producida por las razas de ovinos que hay en el país, el mohair es producido por los caprinos de raza Angora, el cashmere es producido por algunos genotipos de caprinos criollos y las fibras de llama, guanaco y vicuña, son producidas por estos respectivos camélidos sudamericanos.

El comercio internacional de fibras sufre pocas regulaciones y básicamente responde a la oferta y demanda.

Argentina históricamente ha sido muy competitiva en el mercado mundial de lanas, siendo actualmente el cuarto exportador mundial [7]. La competitividad se basa en el volumen que ofrece el país, el bajo costo de producción y la alta calidad. Los bajos costos de producción se deben a la localización de la producción en ambientes de pastizales naturales y manejo extensivo con bajo nivel de insumos.

La calidad de las lanas más finas patagónicas se centra en un muy buen grado de blanco y brillo, pureza, bajos niveles de contaminación vegetal y buena suavidad.

Para el caso del mohair, Argentina es el segundo productor mundial [7], esta fibra tiene buen mercado y se produce en forma competitiva con estándares de calidad, volumen predecible y adecuada descripción.

Para el caso de las fibras de los camélidos silvestres el país tiene grandes oportunidades considerando que es primero en población de guanacos y segundo en vicuñas. En zonas más desfavorecidas, con 3,9 millones de llamas y 3,3 millones de alpacas, la producción total de fibras de camélidos supera los 5 millones de kilogramos anuales. Cerca del 30% de la producción de fibra se transforma y es usada a nivel de predio o comunidad. Alrededor del 80% de la alpaca comercializada es de color blanco y tan sólo el 12% tiene diámetros de fibra menores a los 23 micrones [6].

El valor de la fibra textil está dado, fundamentalmente, por su finura promedio además de otras propiedades que hacen a establecer su cotización tales como el índice de confort PF (*Prickle Factor)* que constituye el porcentaje de fibras con diámetros mayores a 32 micrones, la presencia o ausencia de medulación[1], el crimpado[2] y la

[1] La medulación constituye un canal hueco en el centro de la fibra que supone un problema importante para la industrialización, especialmente en el teñido, porque causa una mayor refracción de la luz haciendo aparecer a las fibras teñidas más claras.

forma y altura de las escamas [8]. Para determinar una medida satisfactoria de calidad de la fibra textil de origen animal, además de tener en cuenta defectos obvios como la pigmentación y la presencia de fibras atípicas o meduladas [9], la característica de mayor importancia es el diámetro medio. Fibras más finas tienen más aplicaciones industriales y en consecuencia tienen mayor valor económico [10].

3 Descripción del problema

Uno de los problemas más importantes que se presenta, en el mercado textil, es poder determinar la distribución del diámetro y la forma de la fibra como parámetro de calidad [11], además de otros factores. En nuestro país existe poca información aún sobre los valores de Coeficiente de Variación de diámetros de fibra (CV) e índice de confort [12] que permita lograr mejoras genéticas por selección y elevar el porcentaje de especímenes con diámetros menores a los 23 micrones.

La evolución de la adopción tecnológica es lenta en los sistemas ganaderos extensivos. Los principales avances esperables a mediano plazo son la especialización y la intensificación de la producción que incluye el uso de nuevas tecnologías y métodos de comercialización más sofisticados. Estos avances se basan en el uso de tecnologías de información y comunicación (TICs) para mejorar la información y capacitación de todos los actores de la cadena. En particular se espera que la comercialización de fibras y de animales progresivamente se base en evaluaciones objetivas y que esa información, junto a la de mercado esté al alcance del productor.

Desde el punto de vista técnico, en la actualidad, se aplica el uso de microscopios de proyección conocidos como *lanámetros* [12] para medir los diámetros de un número determinado de fibras, y a partir de éstos calcular el promedio de diámetros de fibra (PDF) de la muestra analizada, limita su utilización a un número grande de muestras y a una mayor proporción de fibras por muestra.

El desarrollo del *Air Flow* [13] como instrumento de medición rápido y preciso, constituyó un avance importante para generalizar el análisis de muestras de vellones individuales. Con *Air Flow* se obtiene el PDF de la muestra, como resultado de un gran número de fibras, pero nada informa este instrumento de la frecuencia de los distintos diámetros presentes en la muestra [12].

En los últimos años se ha extendido el uso de nuevos instrumentos de determinación de diámetro de fibras como OFDA® (*Optical Fiber Distribution Analyser*) [14] basado en un analizador de imágenes de muestras de fibra y *Sirolan Laserscan*® [15], un lector de fibras por rayos laser. Ambos instrumentos miden en forma rápida y precisa los diámetros de una gran cantidad de fibras, y a través de programas de computación apropiados, grafican la distribución de frecuencia de los diámetros medidos calculando el diámetro promedio. Las muestras utilizadas en estos procesos de medición, se basan en cortes de vellón de aproximadamente 2 mm de longitud y la medida se obtiene a partir de la captura de diámetros longitudinales.

Investigaciones biomecánicas más recientes, demuestran que el análisis del corte transversal provee mediciones más directas y exactas de la finura y madurez de la

[2] El crimpado u ondulado, se refiere a un efecto mecánico producido justamente para lograr cohesión entre fibras iguales. Este factor se relaciona con la capacidad hidrófuga (absorción de humedad) de la fibra.

fibra, usualmente utilizadas para validar y calibrar otras medidas indirectas de estas propiedades esenciales [11]. A pesar de su importancia e interés, los métodos transversales para análisis de imágenes, no se aplican más ampliamente aún a las mediciones de calidad, debido al complejo procesamiento de las imágenes que se obtienen en laboratorio por microscopía de escaneo electrónico (SEM) o por requerir de la intervención de un operador calificado que efectúe manualmente la selección de los diámetros a medir, si se emplea un software como SigmaScan Pro 5.0 para procesar la imagen del corte transversal de la fibra, lo que introduce un considerable error en las mediciones y acarrea la indeseable característica de ser irrepetible.

Sin embargo, la caracterización del corte transversal de la fibra textil atrae considerable interés, ya que el tamaño y forma de las mismas tienen un impacto importante en las propiedades físicas y mecánicas de la fibra [16] cuyas aplicaciones industriales son directas.

En cada medida se tiene que tener en cuenta que dada la gran variación de diámetros que tienen las fibras animales diversas e incluso las vegetales, un gran problema es la exactitud y la precisión. Este es un concepto físico y estadístico respectivamente. Siendo la exactitud la relación entre la medida que hace el aparato y la verdadera medida (en grado de definición en el caso de los microscopios) y la precisión, la repetición, o sea, la relación entre las sucesivas secciones de medida que se pueden hacer (es decir, cómo las medias o promedios de las sucesivas medidas se acercan a las obtenidas previamente). Estadísticamente se determinan también el desvío estándar y el coeficiente de variación de diámetro de las fibras medidas en cada sección [17].

La evolución en los modelos y algoritmos de procesamiento de imágenes en fibras textiles, comienzan con algunos trabajos sobre fibras de algodón que demuestran que las propiedades más relevantes pueden medirse a partir de imágenes microscópicas capturadas en cortes longitudinales y/o transversales. Huang et al. [18] [19] [20] [21] analiza el proceso de medición en el que la imagen de una fibra en corte longitudinal.

En trabajos posteriores, Huang et al. [22] analiza imágenes de fibras de algodón en corte transversal. Mediante este análisis, se aseguran mediciones directas y exactas de la finura y madurez de la fibra. Este método de medición se utiliza, además, como medio de calibración de otros métodos. Para llevar a cabo la medición se recurre a un proceso computacional de segmentación que consiste en la separación de la imagen objeto del resto de los objetos y del fondo. Se utiliza también la técnica de Umbral Adaptativo para preservar el detalle de los bordes y luego, para separar en una primera aproximación los objetos del fondo, se recurre a la técnica de Inundación de Fondo. Finalmente, se implementa un proceso de esqueletizado de la fibra para determinar un punto referencial a partir del cual puedan obtenerse medidas geométricas.

4 Propuesta desarrollada

Como ya se dijo, en nuestro país se tiene poca información sobre valores habituales de coeficiente de variación de diámetro de la fibra y factor de picazón. En ese contexto aparece la necesidad de desarrollar técnicas que permitan medir con la mayor precisión y al menor costo posible, el diámetro promedio de fibra textil para el

análisis y aplicación de metodologías de mejoramiento genético, usos comerciales e industriales.

4.1 Objetivos

La propuesta consiste en el desarrollo de un sistema que permita procesar una imagen del corte transversal de fibra textil de origen animal y proporcione una medida promedio del diámetro de las fibras. El presente trabajo intenta contribuir en el proceso de obtención de medidas de diámetros de fibra confiables, para soportar indicadores de calidad de la fibra. Además, pretende constituirse en una herramienta útil y accesible que dará respaldo a las investigaciones científicas que el SUPPRAD lleva adelante para intervenir en proyectos de Desarrollo y Promoción Humana, y así conducir planes y formular recomendaciones viables para evitar la degradación de los recursos naturales y soslayar problemas de pobreza, marginalidad, emigración y desarraigo entre otros. Finalmente, proporcionar una solución adecuada que permita difundir las cualidades de sustentabilidad para avalar comercialmente los productos textiles de la región.

El aporte fundamental en cuanto a innovación tecnológica radica en el hecho de que los instrumentos actuales de análisis de fibras son costosos y permiten obtener la medida de diámetros en forma longitudinal. En cambio, en el presente trabajo se propone un llevar a cabo un proceso de medición de diámetros en forma transversal con hardware y software de bajos costos, en forma totalmente automatizada y eficaz, mediante el uso combinado de técnicas de procesamiento de imágenes y que puede ser llevada a cabo por personal sin capacitación técnica alguna.

4.2 Descripción

En base a las investigaciones previamente citadas, se desarrolla un sistema que permite obtener una medida del radio promedio de la fibra, a partir del procesamiento automático de la imagen de un corte transversal.

El proceso supone varias etapas en el tratamiento de la imagen para lograr identificar, separar y posteriormente medir la fibra.

Inicialmente, se analizaron las estructuras de diferentes estándares gráficos y se eligió el estándar BMP que consiste en un archivo de mapa de bits con píxeles almacenados en forma de tabla de puntos que administra los colores como colores reales, o bien, usando una paleta indexada. Una de las ventajas de este formato gráfico es que permite obtener un mapa de bits independiente del dispositivo de visualización periférico. Las imágenes se codificaron en 24 bits por píxel, es decir, un byte para cada píxel (16.777.216 colores), color verdadero de alta definición, que se consideró un estándar de fácil manejo desde el punto de vista matemático y de procesamiento y que puede contener la mayor cantidad de información de interés respecto de la imagen original (el uso de 32 bits x pixel sólo agrega efectos de transparencia).

Una vez que la imagen es convertida al formato BMP 24 colores, se la somete a una serie de procesos con el fin de subsanar los defectos que puedan provenir de su captura y para conservar sólo aquellas características que resulten de interés en el proceso de medición:

a) Se convierte la imagen en colores a escala de grises asignando un mismo valor para los bits correspondientes a RGB mediante la aplicación de la expresión obtenida experimentalmente:

$$Y = valor\ R * 0.299 + valor\ G * 0.599 + valor\ B * 0.111$$

Esta expresión se relaciona con la luminancia. La CIE (Comisión Internacional de Iluminación) define la brillantez como el atributo de una sensación visual de acuerdo con el cual un área parece mostrar más o menos luz, siendo la brillantez una cantidad subjetiva, la luminancia ***"Y"*** se considera una forma objetiva de medir la cantidad relacionada con el brillo. Se trata de una función de transferencia no-lineal denominada corrección gamma. En sistemas gráficos por computadora ***"Y"*** es un parámetro numérico que describe la no linealidad de la reproducción de la intensidad, esta codificación maximiza la imagen perceptual.
Por otro lado, al representar un conjunto de colores en tonos de gris, necesitamos manipular sólo 256 valores diferentes.

b) Se ecualiza la imagen construyendo un histograma de frecuencias de grises y se calcula el umbral de binarización, tomando en consideración el tipo de histograma que, en general, resulta asimétrico por dificultades de exposición y foco al momento de la captura, para obtener una imagen con un histograma de distribución más uniforme [22].
c) Se procede a binarizar la imagen en sus valores extremos, tomando como punto binarización el umbral calculado. Con esto se obtiene una imagen en blanco y negro donde se puede distinguir más claramente forma y fondo.

La figura 1 ilustra la anterior secuencia de pasos aplicados al procesamiento de la imagen de un corte transversal de fibra de guanaco y los resultados obtenidos.

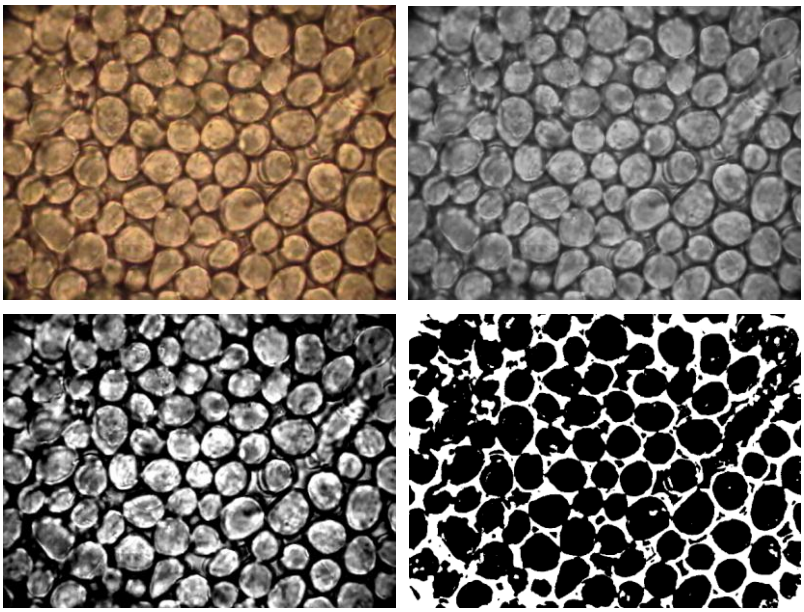


Fig. 1. Imagen superior izquierda: original. Imagen superior derecha: en escala de grises. Imagen inferior izquierda: escala de grises ecualizada. Imagen inferior derecha: binarizada.

Una vez binarizada la imagen, se procede a separar, identificar los objetos a medir.

Para ello se ensayaron varios procedimientos tales como aplicar una convolución de matrices a la imagen según los métodos de Sobel y de Prewit para detectar bordes.

$$\begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & +1 \\ -1 & 0 & +1 \\ -1 & 0 & +1 \end{bmatrix} \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ +1 & +1 & +1 \end{bmatrix}$$

Matrices de Sobel **Matrices de Prewitt**

Como puede verse en la figura 2, este procedimiento no presenta utilidad por obtenerse una imagen en la que no se distingue forma y fondo.

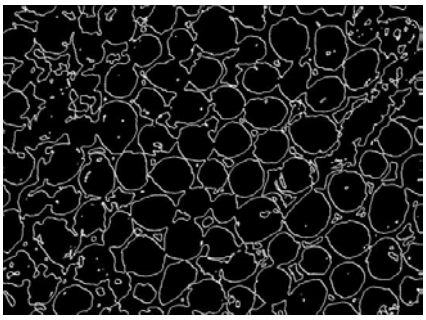


Fig. 2. Resultado de aplicar detección de bordes usando matrices de convolución.

Se ensayo tambien un metodo de adelgazamiento-engrosamiento consistente en inicialmente quitar sucesivas capas a los objetos, de manera tal que los objetos más pequeños (que no serán considerados para la medición y que resultan en "ruido" de la imagen binarizada) sean eliminados. Luego se agregan nuevamente las capas como se ilustra en la figura 3.

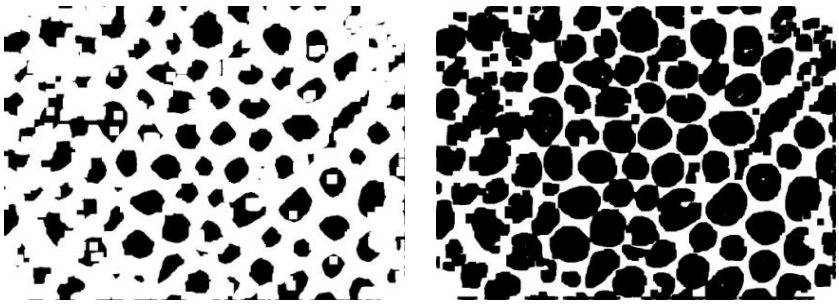


Fig. 3. Imágenes resultantes de aplicar adelgazamiento-engrosamiento

Este método barre el fondo limpiando "el ruido" y se acerca en cierto grado a la separación de objetos, pero presenta la característica indeseable de que se puede perder la forma original del objeto ya que no "guarda memoria" de la forma original.

Por último se ensayó un método de erosión-recuperación, similar al anterior, que presenta la ventaja de conservar la forma original de los objetos porque no elimina las

capas originales y, en cada paso sucesivo, se aproxima al centro geométrico de cada objeto. No obstante, tal como podemos observar en la figura 4, no contribuye a la separación de los objetos de interés ni elimina el ruido de fondo.

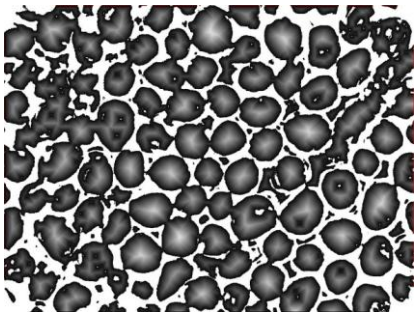


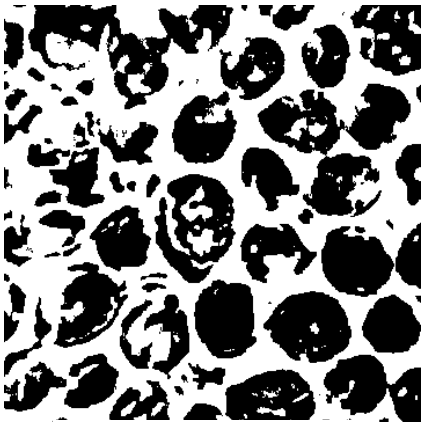
Fig. 4. Imagen resultante de aplicar erosión-recuperación.

Finalmente se implementó una solución que combina algunas de las ideas anteriores con otras nuevas y que proporciona un resultado adecuado. Se aplica un proceso sucesivo de erosión sobre la imagen al mismo tiempo que se conserva en una matriz *m* x *n* del tamaño en píxeles de la imagen original, la información correspondiente a cada nivel de erosión. Los objetos que en este proceso de erosión corresponden a un "adelgazamiento" total en un número de pasos pre-establecido experimentalmente, se rechazan considerándose "ruido" en la imagen. Aquellos objetos cuyo nivel de "adelgazamiento" supera el valor pre-establecido, se consideran de interés y pueden ser recuperados a partir de la información guardada en la matriz de información correspondiente.

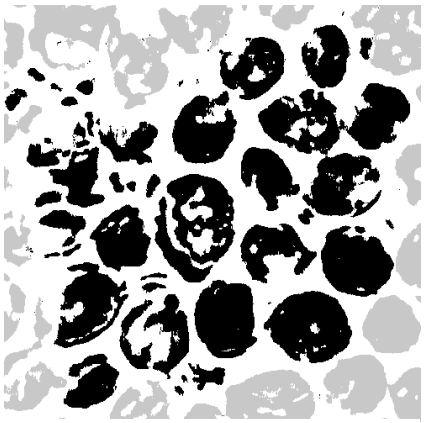
En consecuencia, el proceso completo consiste en sucesivos barridos de la imagen binarizada, en cada uno de los cuales, se descartan primeramente las fibras que se encuentren en contacto con los bordes de la imagen ya que se desconocen sus dimensiones reales, se distingue entre fondo y forma, se rellenan sectores interiores (sólo aquellos sectores que corresponden a componentes conexas y que son interiores al objeto), se seleccionan los objetos a medir tomando en consideración que todo aquello que presente interés en ser medido, no debe exceder ciertos rangos máximo y mínimo entre los cuales puede tratarse de una fibra. Cabe destacar que los valores máximos y mínimos son proporcionados por el laboratorio y corresponden a tamaños esperados en la fibra animal. Por tratarse de fibras naturales, existe un rango apreciable y bastante conocido dentro del cual, el objeto presentado, puede ser considerado o no una fibra.

Por último, se identifica un centro geométrico de dichos objetos tomando como referencia el píxel más "profundo" obtenido en el proceso de erosión-recuperación usado anteriormente y, a partir de éste, se miden 32 radios como distancia a los bordes de la figura. Se calcula el radio promedio y se aproxima la figura a una circunferencia. La aproximación a una circunferencia se debe a que las fibras naturales tienen una geometría muy simple y a los fines prácticos de determinación de calidad esta aproximación es válida y estándar.

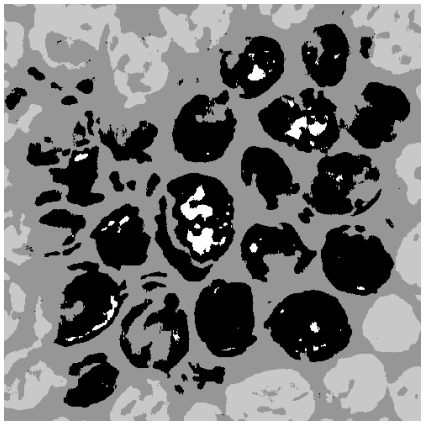
La figura 5 muestra una secuencia de imágenes que ilustran los pasos del proceso detallado en el párrafo anterior.



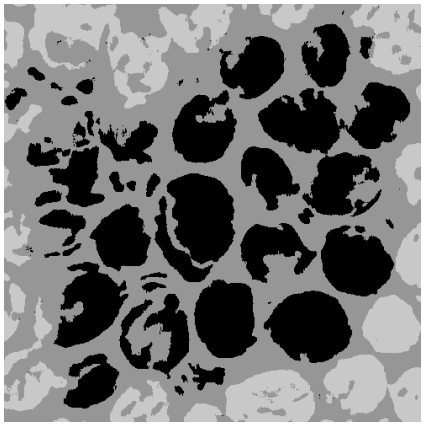
Paso1. Imagen binarizada.



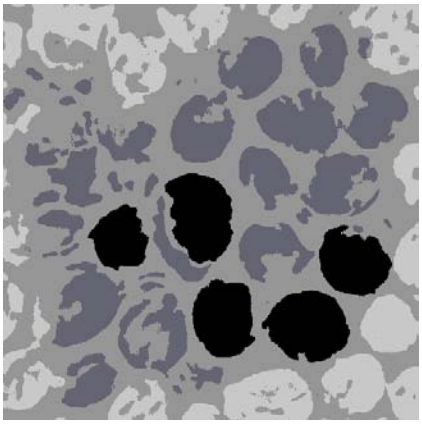
Paso 2. Eliminación de objetos en contacto con los bordes.



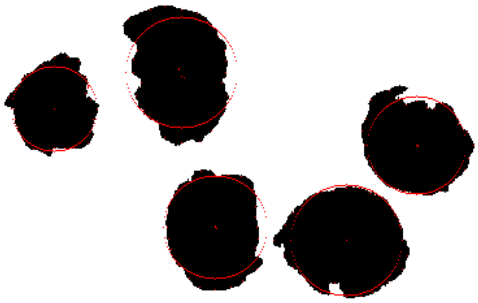
Paso 3. Separación fondo – figura



Paso 4. Relleno de blancos internos.



Paso 5. Selección de objetos útiles a medir.



Paso 6. Determinación del centro geométrico y radios promedio.

Fig. 5. Etapas del proceso automatizado de solución.

4 Discusión de Resultados

El método resultó satisfactorio principalmente porque las mediciones que se obtuvieron en pixeles, con la equivalencia 2 pixeles = 1 micra de acuerdo al aumento del microscopio con el que se capturaron las imágenes, resultaron en valores adecuados para el radio promedio de las fibras, en comparación con los obtenidos a partir de otros métodos de medición en laboratorio y existe total independencia del operador en la medición lo que asegura la precisión y exactitud requeridas. Es decir que el proceso es repetible y su exactitud conocida.

Por otra parte, como ya se dijo, se usan 32 medidas de distancia del centro geométrico de cada objeto a medir para calcular el radio de la fibra ya que experimentalmente se prueba que no se presentan mejoras tangibles en las medidas por aumentar este número.

5 Conclusiones

Por último, y atendiendo la problemática a campo que presenta la determinación de la calidad de la fibra a partir del conocimiento de su finura, el método es útil por cuanto presenta características de buena perfomance, bajo costo de equipamiento y no requiere operación por parte de personal calificado.

Por último, cabe destacar que el software fue desarrollado en Java lo que lo hace portable en cuanto a la plataforma y no involucra costos adicionales en licencias.

6 Trabajo Futuro

En adelante se deben centrar los esfuerzos en la evolución del sistema metrológico atendiendo tres factores principales:

- Revisión de la metodología de captura de imágenes.
- Adecuación de los algoritmos al tratamiento de fibras con otras características morfológicas.
- Reconocimiento de patrones morfológicos a través de redes neuronales.

7 Agradecimientos

Agradecemos al Dr. Eduardo Frank por la generosidad con que nos brindó su tiempo y su conocimiento, que sirvieron de guía en nuestro trabajo.

Referencias

1. Chehébar, C. y Martín S., 1989. Guía para el reconocimiento microscópico de los pelos de los mamíferos de la Patagonia.
2. Lindenmayer, D., Incoll, R., Cunnigham, R. Pope, M. Donnelly, C. McGregor, C., Tribolet, C. y Triggs B., 1999. Comparison of hairtube types for the detection of mammals. Wildlife Research, 26: 745-753.
3. Hausman, L., 1925. A comparative racial study of the structural elements of Human head-hair. The American Naturalist, 59 (665): 529-538.
4. Hausman, L., 1920. The microscopic identification of commercial fur hairs. The scientific Monthly, 10 (1): 70-78.

5. Ford, J. y Roff, W., 1954. Identification of Textile and Related Fibres. J. Textile Inst., 45: 580-611.
6. Programa SUPPRAD. Sustentabilidad Productiva de Pequeños Rumiantes en Áreas Desfavorecidas. UCC, Facultad de Veterinaria.
7. INTA, 2011. Programa Nacional Fibras Animales. Documento Base actualizado a noviembre de 2011.
8. Adot, O., 2010. Introducción a la Industrialización de la Lana y las Fibras Especiales. Documento Interno SUPPRAD N° 2 (2010).
9. Cancio, A., Rebuffi, G., Mueller, J., Duga, L. y Rigalt, F., 2006. Parámetros Cualitativos de la Producción de Fibras de Llamas (Lama Glama) Machos en la Puna Argentina, INTA EEA Bariloche, INTA AER Trancas, INTA EEA Catamarca, Comunicación Técnica, PA 492.
10. Mueller, J., 1993. Objetivos de Mejoramiento Genético para Rumiantes Menores, INTA EEA Bariloche, Comunicación Técninca, PA 238.
11. Frank, E., 2008. Camélidos Sudamericanos. Producción de fibra, bases físicas y genéticas. Revista Argentina de Producción Animal. Vol. 28, pp. 112-119.
12. Mueller, J., (2002). Novedades en la determinación del diámetro de fibras de lana y su relevancia en programas de selección. Comunicación Técnica. INTA, Bariloche, 330pp.
13. Rodriguez Iglesias, R., 1998. Principales características que afectan el valor textil de la lana. Producción Ovina. Dpto. de Agronomía. UNS. Rev. 30/10/07.
14. Qi, K., Lupton, C., Pfeiffer, F. y Minikhiem, D., 1994. Evaluation of the Optical Fibre Diameter Analyser (OFDA) for Measuring Fiber Diameter Parameters of Sheep an Goats. Jounal Animal Sci. 72: 1675-1679.
15. Guzmán Barzola, J.C. y Aliaga Gutiérrez, J.L., 2010. Evaluación del Método de calsificación del Vellón en Ovino Corriedale (Ovis Aries) en la Sais Pachacutec. Producción Animal. Facultad de Zootecnia UNALM.
16. Xu, B., Pourdeyhimi, B. y Sobus, J., 1993. Fiber Cross-Sectional Shape Analysis Using Image Processing Techniques, Textile Research Journal, Vol. 63, N° 12.
17. Frank, E., Hick M., Prieto, A., Castillo, M., 2009. Metodología de Identificación Cualitativa y Cuantitativa de Fibras Textiles Naturales. Documento Interno SUPPRAD N° 1 (2009).
18. Huang, Y. y Xu, B., 2002. Image Analysis for Cotton Fibers. Part I: Longitudinal Measurements. Textile Research Journal, 72(8), 713-720.
19. Xu, B. y Ting, Y., 1996. Fiber Image Analysis. Part I: Fiber Image Enhancement. Textile Research Journal, 87, 274-283.
20. Xu, B. y Ting, Y., 1996. Fiber Image Analysis. Part II: Measurement of General Geometric Properties of Fibers. Textile Research Journal, 87, 284-295.
21. Rojas Vigo, D. A., 2006. Caracterización del Espesor de las Fibras de Alpaca Basada en Análisis Digital de Imágenes. Electrónica-UNMSM, N° 17.
22. Huang, Y. y Xu, B., 2004. Image Analysis for Cotton Fibers. Part II: Cross-Sectional Measurements. Textile Research Journal, 74(5), 409-416.

Nuevos descriptores para la identificación de personas basados en la simetría del trazo

Verónica I. Aubin (1), Jorge H. Doorn (1,2), Gladys N. Kaplan (1)

(1) Departamento de Ingeniería e Investigaciones Tecnológicas
Universidad Nacional de La Matanza, Florencio Varela 1903, San Justo, Argentina.
(2) INTIA, Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Provincia de Buenos Aires,
Paraje Arroyo Seco, Campus Universitario, Tandil, Argentina.
e-mail: vaubin@unlam.edu.ar, jdoorn@exa.unicen.edu.ar

Resumen. La identificación de autores de trazos manuscritos es un área del procesamiento de imágenes en la que se han realizado muchos aportes en los últimos años. Sin embargo se trata de un dominio en el que restan aspectos por resolver. En el presente trabajo se reportan resultados relacionados con la identificación de nuevos parámetros descriptores que contribuyen a la identificación del autor del trazo. Estos parámetros se obtienen de los residuos observables en el papel luego de realizar un escrito.

Keywords: grafología, análisis de trazos, presión del trazo, perfiles residuales.

1 Introducción

Tal como ocurre con la mayoría de los parámetros vitales de un ser humano, los trazos manuscritos realizados por una persona tienen una gran variabilidad dependiendo de numerosos factores. Sin embargo en los trazos manuscritos existe un núcleo de aspectos invariantes que hace viable el reconocimiento manual o automático del autor.

Solange Pellat hijo, fue uno de los principales constructores de la grafonomía, junto con el filosofo André Lalande. Desde 1903 Pellat demostró y dio origen a un análisis fino de los movimientos de la escritura, realizando una investigación sobre las leyes fundamentales de la escritura. Finalmente en 1927 publicó una síntesis de sus trabajos en donde enuncia una serie de leyes que actualmente se utilizan. [1]. Citando a Viñals y Puente "La fiabilidad del sistema era fruto de una nueva visión sobre la escritura: [2] no considerarla como arte, sino como un reflejo fisiológico y psicológico del individuo. Es por ello que se convierte en un elemento identificativo <…>. Pero en la escritura existe una jerarquía de signos, tales como la profundidad, intensidad, presión, rapidez, dirección, continuidad que son muy difíciles de imitar. Esta metodología que se demostró altamente efectiva pues se adentra en la anatomía de la letra, y consiguió el reconocimiento de la Justicia" [2].

En el presente artículo se ha tratado profundizar el abordaje analítico del grafismo, utilizando recursos del procesamiento de imágenes. Procurando encontrar aspectos característicos del trazo, invisibles o muy poco visibles, que tengan la propiedad de ser altamente repetitivos.
El estudio semi-automático o automático de trazos manuscritos es una actividad de importancia en una gran variedad de dominios. En algunos de ellos el objeto de interés es la comprensión del texto escrito, mientras que en otros se procura identificar al autor del mismo [3]. Este artículo se concentra en algunos aspectos relacionados con la identificación del autor de los trazos.
La fuerza ejercida durante la escritura debe medirse durante el acto de la escritura propiamente dicha. Existen dispositivos [4] [5] [6] que permiten conocer esta fuerza y los estudios realizados con los mismos han sido valiosos para estimar las características de las mismas. Sin embargo son poco transportables al problema práctico ya que en general los estudios de identificación del escribiente sólo tienen acceso al resultado de la escritura.
La fuerza ejercida en el momento de la escritura deja algunos residuos tales como el color relativo de cada fragmento del trazo o el ancho del mismo. Este hecho ha sido reconocido y analizado por muchos autores [7] [8] [9] [10].
Los trabajos cuyos resultados se reportan en este artículo comenzaron analizando algunos aspectos de los trazos manuscritos, a través del procesamiento de imágenes. Se relacionó, bajo condiciones controladas, la fuerza ejercida cuando una persona escribe, con el grosor y valor de gris del trazo [11] [12] [13].
Se estableció que el ancho medio y el valor de gris son, dentro de ciertos límites, casi proporcionales al peso, pero una vez que el papel alcanza la máxima deformación condicionada por la base ya no varía significativamente. Se comprobó que no había variación en los resultados anteriores utilizando distintos colores de tinta [11] [12].
Además, se encontró que un trazo espontáneo de un grafema aparecen zonas donde el ancho medio y el valor de gris son notoriamente diferentes del resto del trazo. Estas zonas son casi invariantes en su ubicación relativa para todas las muestras del mismo grafema realizadas por la misma persona [13].
Las comparaciones realizadas sobre los gráficos característicos del trazo arrojaron resultados muy favorables. Por un lado los valores de grises y los anchos medios del trazo son altamente repetitivos para trazos que representan el mismo grafema realizados por el mismo autor [13].

2 Identificación del autor mediante niveles de grises

El término “pseudo-dinámicas” se usa para distinguir datos reales dinámicos, grabados durante el proceso de escritura, de información, que puede ser reconstruido de la imagen estática. En la literatura se han propuesto diversas características de este tipo, las cuales se revisan a continuación. Una gran cantidad de los trabajos publicados en este campo tratan la autentificación de firmas, la cual es un subproblema de la identificación del autor desde un texto cualquiera.

En [7] se analiza el histograma de la imagen en escala de grises de la firma, y se propone el cómputo de un umbral de presión (Umbral de Alta Presión). En conjunto

con dicho parámetro, se representa la firma mediante el porcentaje de pixeles que superan el umbral, el valor mínimo y máximo de nivel de gris de la imagen, el rango dinámico de la firma, entre otros. Para validar este conjunto de descriptores el trabajo adopta un clasificador basado en distancia. En [8] se propone un método para la verificación de firmas off-line basado en características geométricas y redes neuronales. Las características consideradas son el esqueleto del trazo, el contorno, y regiones de alta presión.

En [14] a partir del esqueleto de la imagen, se detectan los trazos que superen una longitud mínima, y se procede a calcular un Índice de Suavidad (SMI) para cada uno de los trazos. El descriptor corresponde a la razón entre el número de trazos suaves y el número total de trazos de la firma.

En [9] se introduce un enfoque orientado a la detección de falsificaciones elaboradas. Considera información dinámica como la presión que se ejerce con la punta del instrumento de escritura, y pone énfasis en la extracción de los pixeles de baja presión de acuerdo al nivel de gris en la imagen. No se hace referencia en éste trabajo al tipo o gramaje del papel utilizado en las muestras, ni se hace referencia al tipo de instrumento de escritura ni de tinta utilizado, tampoco se menciona la base de apoyo. Los autores utilizan un umbral de decisión basados en una hipótesis de trabajo de que existe una relación entre la presión ejercida en la escritura y valor del nivel de gris en la imagen, pero nunca la probaron.

En [15] el modelo propuesto utiliza un conjunto de características estáticas y pseudo-dinámicas para la verificación. Las características estáticas corresponden básicamente al calibre, el cual describe la relación entre la altura y el ancho, un parámetro que representa la suavidad de la firma, el espaciamiento y la alineación con respecto a una la línea base. Las características pseudo-dinámicas consideradas son la progresión, la forma y la inclinación.

En [16] se propone un método de verificación de firmas Chinas off-line, utilizan tanto características estáticas como dinámicas, y Support Vector Machine para la clasificación. El sistema de verificación propuesto combina cuatro conjuntos de características: Características de momento (la proporción entre la altura y el ancho, el grado de inclinación, el grado de extensión, el grado de excursión horizontal y vertical), distribuciones de dirección, distribución del nivel de gris, y distribución del ancho de trazo. Tampoco los autores hacen referencia al tipo o gramaje del papel utilizado en las muestras ni al tipo de instrumento ni de tinta utilizado, tampoco se menciona la base de apoyo. Para los autores la densidad o presión aparente se describe por la anchura de los trazos, esto es una hipótesis de trabajo que tampoco fue nunca probada.

En [17] se propone un método basado en dos imágenes y su transformación a coordenadas polares. La primera imagen contiene información de los puntos de alta presión, y la segunda corresponde a la versión binarizada de la imagen original. El espacio polar se divide en secciones angulares y radiales, en donde se determina la distribución de los puntos de alta presión. Además, se considera la densidad de dichos puntos respecto del centro geométrico de la firma original. Para poder determinar los puntos de alta presión, se calcula un umbral de alta sobre el histograma de la imagen en escala de grises.

En [10] se propone un método que combina el análisis global y local de la imagen en conjunto con un Support Vector Machine para la clasificación de los patrones. El

análisis global se realiza mediante la estimación de variaciones de nivel de gris de la imagen usando la transformación wavelet, y tratamiento local considera la obtención de información de textura proveniente de la matriz de coocurrencia.

3 Simetría del trazo

Una de las contribuciones principales de este artículo está relacionada con el análisis de la línea que une los puntos más oscuros del trazo. Dado que la línea de mínimos se ubica siempre en la misma posición relativa en las diferentes zonas de un grafema producido por el mismo autor. Estudiando más detalladamente este fenómeno se encontró que la distancia relativa entre la línea de los mínimos y el esqueleto es repetitiva para una persona y varía de individuo a individuo.

El camino que permitió construir la contribución anterior comenzó con la verificación de la hipótesis de trabajo en la que se basan distintos autores, en el sentido que existe una relación entre la presión ejercida en la escritura y valor del nivel de gris en la imagen y el ancho del trazo.

Utilizando un dispositivo diseñado especialmente, se realizaron numerosos trazos rectos con diferentes pesas de manera de realizar trazos con presión constante. Las pesas utilizadas fueron seleccionadas entre 10g y 200g (lo que se corresponde con los 0,1 y 2 Newton).

En la figura 1, se presenta un gráfico de los valores de la escala de grises a lo largo de una línea imaginaria perpendicular al trazo, en un punto determinado del mismo. La zona horizontal representa el papel sin ninguna escritura. Se muestra en la imagen como se tomaron las mediciones de ancho medio del trazo y valor del nivel de gris, para su posterior análisis.

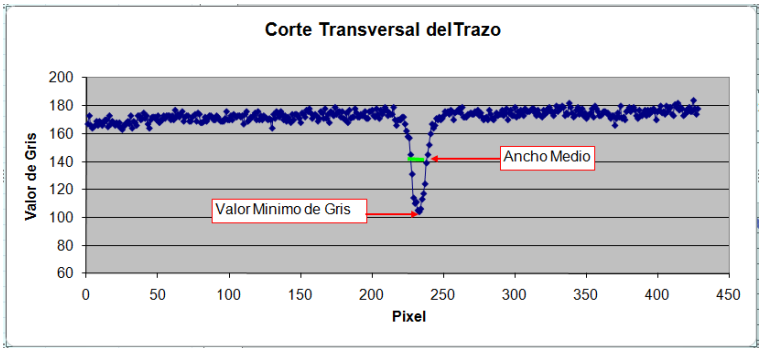
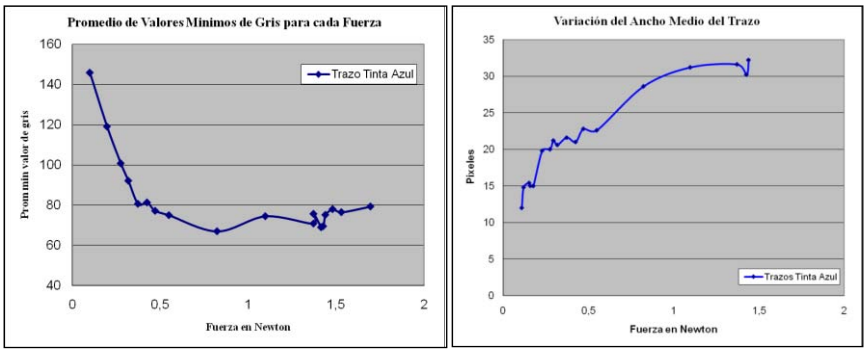


Figura 1 Valor de gris a lo largo de una línea perpendicular al trazo.

Debe notarse que los valores de la ordenada se corresponden con intensidad de tonos de gris, no de fuerza.

La figura 2 muestra lo la influencia en las mediciones de la fuerza aplicada en la escritura. La figura 2 (a) muestra la relación de las distintas fuerzas aplicadas y el valor de gris en el centro del trazo. En la figura 2 (b) se ve la influencia de la fuerza aplicada sobre el ancho medio del trazo.



(a) (b)

Figura 2. (a)Nivel de gris en el centro del trazo (b) Ancho medio del trazo.

Los resultados obtenidos representan una realidad que se ajusta a lo que se esperaba, se observa que entre la cota inferior de fuerza y aproximadamente 1 N el ancho medio y el valor de gris son casi proporcional al peso, pero una vez que el papel alcanza la máxima deformación condicionada por la base ya no sigue adelante.

Se repitió la experiencia para estudiar como influía el color de la tinta en las mediciones. Se observó que se mantienen los resultados obtenidos anteriormente independientemente del color de la tinta utilizada.

3. 1 Metodología propuesta

El arreglo experimental se basa en la captura de imágenes, usando la luz difusa ambiente; el mismo instrumento de escritura, bolígrafo “bic trazo grueso” de color azul y las características del papel y la base de apoyo en 5 hojas de 75g/m2.

El proceso comienza con la umbralización de la imagen y el suavizado de los bordes aplicando los algoritmos de erosión y dilatación [18]. Luego, se esqueletiza el trazo manteniendo la continuidad del mismo [19], y se calcula para cada punto del esqueleto la recta perpendicular al mismo, sobre la cual se mide en la imagen original el valor del mínimo gris. Se procede a rectificar el trazo de manera de poder graficar el valor de gris en un sistema cartesiano.

3.2 Distancia

Como ejemplo, se considera la letra "e”, la cual permite observar con facilidad la composición del descriptor. Se definen segmentos sobre el trazo completo del grafema. Estos segmentos procuran contemplar las zonas relativamente homogéneas del trazo. En la figura 3 se muestran los 7 segmentos en que se dividió el grafema para su análisis, los cuales se encuentran etiquetados con letras desde la a hasta la g.

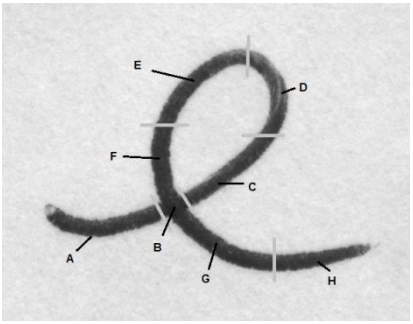
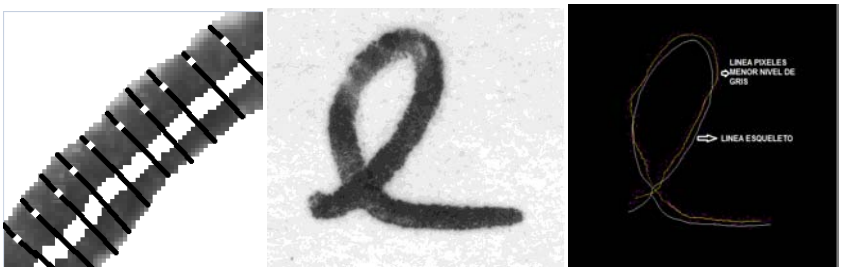


Figura 3 Segmentos del grafema

El proceso de obtención del descriptor es el siguiente, a partir de las líneas perpendiculares de cada uno de los puntos del esqueleto, se identifican las coordenadas y el nivel de valor gris del pixel más oscuro sobre la perpendicular. Los puntos blancos de la figura 4(a) muestra esquemáticamente los pixeles de menor nivel de gris. El cálculo de la distancia entre los puntos del esqueleto y su correspondiente punto más oscuro se realiza sobre la perpendicular la que se señala con los trazos negros en el esquema. Para observar el resultado del computo del esqueleto y la línea de menores niveles de grises en una imagen real, se presenta la figura 4(b) y figura 4(c).

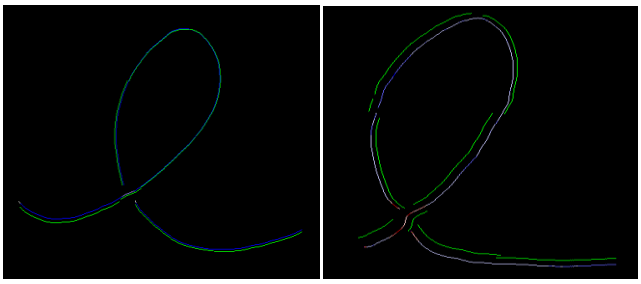


(a) (b) (c)

Figura 4 Proceso de Computo del Descriptor

Por último, para cada segmento del trazo se obtiene el promedio de las distancias. A estos promedios por segmento se les denomina ρP. Considerando entonces que para este grafema se divide en 7 segmentos, el descriptor del grafema completo corresponde al vector constituido por los 7 promedios de las distancias normalizadas en cada segmento. Así, el descriptor considerado corresponde al vector $D = (\rho P_1, \rho P_2, \rho P_3, \rho P_4, \rho P_5, \rho P_6, \rho P_7)$.

En la figura 5 se compara la distancia promedio de dos personas distintas. Se puede observar en la figura 5(a) que la línea que representa el promedio de las distancias esta cerca de la línea del esqueleto para todos los tramos del trazo. En la figura 5(b) la línea del promedio de las distancias tiene una separación mayor con respecto a la línea central del trazo.



(a) (b)

Figura 5. Distancia promedio de dos personas distintas

En los casos que los intervalos de distancia promedio son independientes, se puede utilizar el mismo como característica para clasificar a los autores de las muestras. Evaluando si la distancia de la muestra de la que se quiere probar su autoría se encuentra dentro del intervalo correspondiente.

A modo de ejemplo la figura 6 muestra la distancia ± el error para los segmentos del trazo inicial, cruce, Trazo Ascendente Inferior, Trazo Ascendente Superior, Trazo Descendente Superior, Trazo Descendente Inferior, Trazo Final 1ra Mitad y Trazo Final 2da Mitad del grafema 'e' realizado por dos autores distintos.

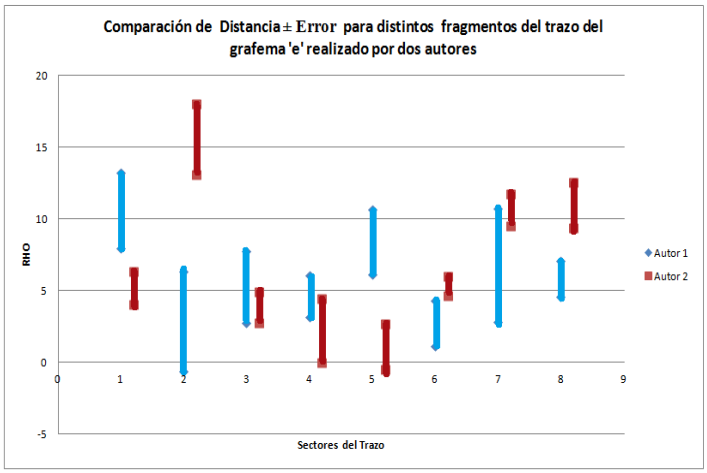


Figura 6 Comparación de la distancia promedio ± Error, del grafema realizado por dos autores.

3.3 Puntos de Cruces

Se analizaron los cruces de la línea que une los puntos de mínimo valor de gris con la línea del esqueleto del grafema.

En la figura 7 se observan los puntos de cruce y el sentido de los mismos en relación con la línea del esqueleto. Las figuras 7(a), 7(b) y 7(c) corresponden a

muestras realizadas por la misma personal, se visualiza en ellas que las zonas y los sentidos de los cruces se repiten. El cruce en el trazo ascendente es de adentro hacia afuera del trazo. Cruce en el trazo descendente es de afuera hacia adentro del trazo. Para la ubicación del cruce en el grafema se decidió utilizar un porcentaje del total del tramo del trazo en el que se encuentra. Por ejemplo, en la figura 7(a), el cruce 1 se ubica al 66% del tramo ascendente y el cruce 2 a 60% del tramo descendente.

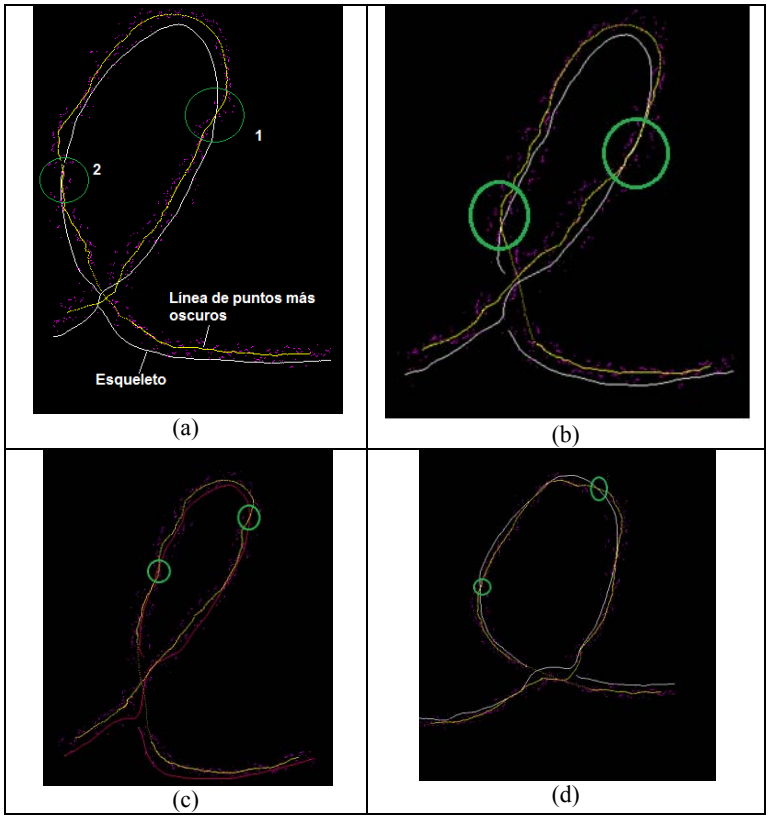


Figura 7 Puntos de Cruces

Para otra persona, se puede observar en la figura 7(d) los cruces. Cruce en el trazo ascendente es de afuera hacia adentro del trazo ubicado al 85% del trazo. El cruce en el trazo descendente es de adentro hacia afuera del trazo al 60% del trazo, como se observa en la Tabla1.

Tabla 1Ubicación de los cruces en cada grafema medidos en porcentajes del tramo.

	Persona1			Persona2
Cruce	Imagen 16	Imagen 17	Imagen 18	Imagen 19
Tramo Ascendente	66%	60%	75%	85%
Tramo Descendente	60%	65%	55%	60%

4 Conclusiones

En este artículo se presenta un método no invasivos de bajo costo que permiten extraer características del trazo manuscrito ya producido. Analizando las diminutas deformaciones que la escritura produce sobre el papel y las características del trazo tales como valor de gris del mismo.

El método propuesto no modifica física o químicamente el texto original, lo que posibilita múltiples análisis. Esta característica lo hace muy atractivo para ser utilizado en análisis forenses, ya que permite preservar la muestra original.

Se confirmaron y ampliaron los resultados de otros autores en el sentido que los valores de gris [7] [9] [17] y los anchos medios del trazo [16] son altamente repetitivos para trazos que representan el mismo grafema realizados por el mismo autor.

En este artículo se planteó la extracción de características en función del valor del nivel de gris a lo largo del trazo, en lugar de utilizar histogramas como plantearon otros autores [7] [9] [17]. Un mismo histograma puede responder a trazos con diferentes características, mientras que en el método propuesto el valor de gris está relacionado con su ubicación en el trazo.

Se ha comprobado que la relación de la línea que une los puntos más oscuros del trazo con el esqueleto brinda información muy útil para identificar al autor. Los parámetros que se pueden extraer de esta comparación son:

- Distancia de separación en pixeles.
- Los cruces y su ubicación en el trazo.
- Posiciones relativas con respecto al esqueleto.

Se han confirmado las conclusiones de muchos autores, en el sentido que no hay una única característica del trazo que se pueda considerar suficiente para identificar el autor del trazo. Los resultados de este trabajo permitieron ampliar el conjunto de características pseudo-dinámicas.

Trabajos Futuros

En virtud que todos los estudios fueron realizados con una cantidad exigua de muestras, se hace necesario ampliar la cantidad de muestras en al menos tres dimensiones:

- Más muestras de cada grafema producida por el mismo autor.
- Más autores.
- Más grafemas.

Queda también pendiente evaluar el impacto de la inclinación del instrumento de escritura sobre la simetría de los valores de gris en la dirección perpendicular al esqueleto del trazo. Potencialmente puede ser necesario regresar a la realización de trazos controlados pero con el instrumento de escritura en diferentes ángulos respecto del papel.

Los estudios de grosor y el análisis de la presencia de “minucias” serán reportados en futuros artículos.

5 Referencias

1. Manuel José Moreno Ferrero “Grafología Forense: La Pericia Caligráfica Judicial” www.grafoanalisis.com/moreno_forense.pdf (última consulta marzo 2011)
2. F. Viñals and M. Puente. “Pericia Caligráfica Judicial: Práctica, casos y modelos”. Ed. Herder, Barcelona. 2001.
3. Plamondon R. y Srihari S. N “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey”. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 63–84 (Jan. 2000).
4. D. Sakamoto, T. Ohishi, Y. Komiya, H. Morita and T Matsumoto, "On-line Signature Verification Algorithm Incorporating Pen Position, Pen Pressure and Pen Inclination Trajectories", Proc. IEEE ICASSP 2001, Vol. 2,pp. 993-996, 2001.
5. W. S.Wijesoma, K.W. Yue, K. L. Chien, and T. K. Chow, “Online handwritten signature verification for electronic commerce over the internet,”,WI 2001. N. Zhong etal. Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 227–236.
6. Nelson, W. and E. Kishon.. Use of Dynamic Features for Signature Verification. IEEE Transactions on Systems, Man and Cybernetics. 201-205. 1991
7. Ammar, M., Yoshida, Y., Fukumura, T.: “A New Effective Approach for Off-line Verification of Signatures by Using Pressure Features”, Proc. ICPR (1986) 566-569
8. Huang, K., Yan, H.: “Off-line signature verification based on geometric feature extraction and neural network classification”, Pattern Recognition 30(1997) 9-17
9. A. Mitra, P. Kumar Banerjee and C. Ardil. “Automatic Authentication of Handwritten Documents via Low Density Pixel Measurements”. International Journal of Information and Mathematical Sciences 2:4 2006.
10. Vargas, J.F., Ferrer, M.A., Travieso, C.M., Alonso, J.B.: “Off-line signature verification based on grey level information using Texture features”. 2010.
11. V. I. Aubin, R. S. Wainschenker, J.H. Doorn.: “Determinación de Propiedades de Trazos Manuscritos por Medios Interferométricos”. WICC-2005. 2005 ISBN:950-665-337-2 pág 134-137.
12. V. I. Aubin, R. S. Wainschenker, J.H. Doorn.: “Perfilometía Virtual en Trazos Manuscritos Residuales”. WICC 2010.
13. V. I. Aubin, R. S. Wainschenker, J. H. Doorn.: “Aspectos Invariantes en Trazos Manuscritos”. WICC 2011.
14. Fang, B., Wang, Y.Y., Leung, CH., Tang, Y.Y., Tse K.W., Kwok, P.C.K. and Wong, Y.K.: “A smoothness index based approach for off-line signature verification”. In Proceedings of the Fifth International Conference on Document Analysis and Recognition, pages 785–787, 1999.
15. Oliveira L.S., Justino, E., Freitas, C. and Sabourin, R.: “The graphology applied to signature verification”. In 12th Conference of the International Graphonomics Society, pages 286–290, 2005.
16. Lv, H., Wang, W., Wang, C. and Zhuo, Q. “Off-line Chinese Signature Verification Based on Support Vector Machine”. Pattern Recognition Letters, Elsevier, 26:2390–2399, 2005.
17. Vargas, J.F., Ferrer, M.A., Travieso, C.M. and Alonso, J.B.: “Off-line signature verification based on high pressure polar distribution”. In Proceedings International Conference on Frontiers in Handwriting Recognition 2008. Montereal., August 2008.
18. Baxes G. A. “Digital Image Processing” John Wiley & Sons Inc.(1994).
19. Zhang T. Y. and Suen C. Y. “A fast parallel algorithm for thinning digital patterns”. In Communications of the ACM, volume 27, pages 236-239, 1984

XI WORKSHOP TECNOLOGÍA INFORMÁTICA APLICADA EN EDUCACIÓN

- WTIAE -

CACIC2013

XI WORKSHOP TECNOLOGÍA INFORMÁTICA APLICADA EN EDUCACIÓN - WTIAE -

ID	Trabajo	Autores
5712	**Moodle en la enseñanza universitaria: uso novedoso de la actividad libro**	Carina Fracchia (UNCOMA), Ana Alonso (UNCOMA)
5814	**Uso de Software Interactivo como facilitador para la Introducción Temprana de Conceptos de Control Robustos**	Patricia Baldini (UNS), Guillermo Calandrini (UNS), Pedro Doñate (UNS), Héctor Bambill (UTN)
5744	**TICs para una Educación Inclusiva**	Emilce Castillo, Rossana Sosa Zitto, Ulises Rapallini, Rafael Blanc, Leandro Lepratte (UADER)
5636	**Estrategias de aprendizaje en procesos mediados por TIC: una experiencia con alumnos ingresantes**	Tatiana Gibelli (UNRN)
5802	**Epistemological obstacles in the learning process of Numeral Systems**	Marcia Ivonne Mac Gaul, Ma Laura Massé Palermo, Paola Del Olmo (UNSa)
5615	**Avances en el diseño de una herramienta de autor para la creación de actividades basadas en realidad aumentada**	Lucrecia Moralejo, Cecilia Sanz, Patricia Pesado (UNLP), Sandra Baldassarri
5855	**Una plataforma de edicto de aulas acessíveis para profesores: transformando aula em diversidade**	Cristiani de Oliveira Dias, Eliseo Berni Reategui (UFRJ)
5876	**Aprendizaje Basado en Competencias y Objetos de Aprendizaje**	Silvina Bramati, Zulema Rosanigo, Claudia Lopez, Pedro Bramati (UNPSJB)

XI WORKSHOP TECNOLOGÍA INFORMÁTICA APLICADA EN EDUCACIÓN - WTIAE -

ID	Trabajo	Autores
5891	**Animaliƈ: Material educativo digital para la enseñanza de la Zoología**	Sabrina Martorelli, Sergio Martorelli, Cecilia Sanz (UNLP)
5803	**Diseño de una aplicación de Aprendizaje Matemático Basada en Tecnología Android**	Ruben Cáceres, Roy Genoff, Leandro Ayala, Patricia Zachman (UNCAus).
5619	**Primeros pasos en el desarrollo de ambientes virtuales inmersivos de aprendizaje utilizando software libre**	Iris Sattolo, Guillermo Sutz, Hernan Monti, José García (UM), Liliana Lipera
5651	**O Letramento E O ensino De Literatura Mediados Por Jogos Digitais Educacionais**	André Noronha Furtado de Mendonça,Denise Mallmann Vallerius (IF-SUL)
5746	**Una Aplicación Móvil para el Museo de Física de la Universidad Nacional de La Plata**	Javier Diaz, Ivana Harari, Andrea Gallego, Leandro Aguilar (UNLP)
5788	**Implementatión of a 3D Virtual Environment at the National University of the North West of the Province of Buenos Aires**	Hugo Ramón, Claudia Russo, Leonardo Esnaola, Nicolas Alonso, Maximiliano Fochi, Franco Padovani (UNNOBA)
5809	**Web Authoring Tool and Repository for Leaning Objects**	Lucas Ferrari da Costa, Maximiliano Reidel, Vinícius de Carli, Júlia Marques Carvalho da Silva (IF-SUL)
5871	**Incorporar Actividades Virtuales en Educación Superior: Modelo para Caracterizar a los Docentes según sus Competencias**	Lucía Malbernat (CAECE)

XI WORKSHOP TECNOLOGÍA INFORMÁTICA APLICADA EN EDUCACIÓN - WTIAE -

ID	Trabajo	Autores
5884	**Interoperability in virtual World**	Leandro Rosniak Tibola (OTRA), Liane Margarida Rockenbach Tarouco (UFRGS)
5896	**Tecnología informática aplicada a la educación de adultos mayores**	Beatriz Depetriz (UNPSJB), Guillermo Feierherd, Marcela Jerez (UNTDF)
5598	**Moderación de sesiones colaborativas a través de la virtualización de la técnica Metaplan**	Alejandro Gonzalez, Ma. Cristina Madoz, Ma Florencia Saadi, Dan Hughes (UNLP)
5734	**Taxonomía de Mecanismos de Awarenses**	Alexander Herrera (UTN-FRBA), Darío Rodriguez, Ramon García-Martinez (UNLA)
5614	**Web ECALEAD: diseño de un prototipo web como herramienta de soporte para la Evaluación de Calidad en Educación a distancia**	Gladys Gorga, Cecilia Sanz, Ma. Cristina Madoz (UNLP)
5757	**Juegos Educativos Móviles: Aspectos involucrados**	Alejandra Lliteras, Cecilia Challiol, Silvia Gordillo (UNLP)
5610	**Characterizacion of University- Drop –Out at UNR. Using data Mining. A Stydy Case**	Sonia Formia (UNRN), Laura Lanzarini, Waldo Hasperué (UNLP)
5886	**Arte y TIC: Experiencias iniciales con herramientas de software en la formación de Licenciados en Artes Combinadas**	Mirta Fernandez, Walter Barrios, G. Gendin (UNNE), María Viviana Godoy (UNNE)

XI WORKSHOP TECNOLOGÍA INFORMÁTICA APLICADA EN EDUCACIÓN - WTIAE -

ID	Trabajo	Autores
5611	**Aplicación del aprendizaje basado en problemas y la tecnóloga informática a la enseñanza de programación en los primeros años de ingeniería**	Ricardo Coppo, German Feres, Gustavo Ursua, Javier Iparraguirre, Ana Cavallo (UTN- FRBB)
5776	**Construcción de Clasificadores Automáticos de habilidades de E-tutores de Aprendizaje Colaborativo**	Pablo Santana Mansilla, Rosanna Costaguta, Daniela Missio (UNSE)
5826	**Manifestación de habilidades de colaboración en grupos de aprendizaje sincrónico y asincrónico**	Diego Yanacon Atia, Costaguta Rosanna (UNSE)

Moodle en la enseñanza universitaria: uso novedoso de la actividad libro.

Carina Fracchia[1], Ana Alonso de Armiño[1]

[1] Facultad de Informática, Universidad Nacional del Comahue,
Buenos Aires 1400. Neuquén, Argentina
{carina.fracchia, ana.alonso }@fai.uncoma.edu.ar

Resumen. Los avances tecnológicos traen aparejados cambios en la forma de enseñar y aprender. Desde el año 2004 se han realizado diversas experiencias educativas en nuestra Institución que han permitido observar, que una progresiva integración de las nuevas Tecnologías de la Información y la Comunicación (TIC) a los procesos formales de enseñanza y aprendizaje, puede afectar positivamente los aprendizajes, capacidades y habilidades de nuestros estudiantes. En este trabajo se presenta una experiencia realizada en una materia de primer año, donde se ha utilizado la actividad libro en Moodle como una bitácora, registrando lo acontecido cronológicamente en las clases teóricas y prácticas. Este enfoque dado al trabajo con la actividad libro ha permitido presentar la información y material educativo de manera intuitiva y sencilla, además de fomentar y fortalecer el trabajo colaborativo entre los integrantes del equipo docente de la materia donde se llevó a cabo la experiencia.

Palabras claves: TIC, Trabajo Colaborativo, Entornos Virtuales de Enseñanza y Aprendizaje

1 Introducción

En la actualidad la incorporación de las TIC en el ámbito educativo es una realidad en muchas instituciones. Una progresiva integración a los procesos formales de enseñanza y aprendizaje permitirá modificar las prácticas educativas en el seno de las aulas y afectar positivamente, los aprendizajes, capacidades y habilidades de nuestros estudiantes.

Al evaluar el uso de los recursos TIC empleados en nuestra práctica docente podremos identificar los procesos y prácticas que resulten ser más eficaces. Además, será necesario investigar la existencia de herramientas que resulten novedosas y dispositivos que nos ayuden a desplegar estrategias que faciliten ayudar a los alumnos

a aprender. Tal como mencionan los autores Bustos y Román "en otros términos, parece fundamental centrar nuestra atención en los procedimientos, estrategias, mecanismos, dispositivos y experiencias cuyo objetivo es la evaluación de los usos de las TIC para impulsar nuevas formas de aprender y enseñar, a partir de sus hallazgos y resultados". [1]

1.1 Entornos Virtuales de Enseñanza y Aprendizaje

Una de las herramientas que ha cobrado mayor difusión en los últimos tiempos, son las Plataformas para Educación a Distancia, también conocidos como Entornos Virtuales de Enseñanza y Aprendizaje (EVEA). En relación a este tipo de herramientas se han originado un gran número de trabajos de investigación y publicaciones. Plaza [2] hace hincapié en que al decidir hacer un cambio metodológico e incorporar la tecnología en la clase, es fundamental que "el docente lo considere necesario y esté dispuesto a hacer un cambio en la forma de enseñar, en su forma de organizar su materia. Debe considerar a las TICs como herramientas para hacer, mostrar, pensar, sin hacer de ellas el centro de la clase. Pero la inclusión de la tecnología en el aula no debe ser sólo el esfuerzo de uno. El docente debe sentir el respaldo institucional para el cambio ".

Si bien estos entornos de aprendizaje ponen a disposición del estudiante una amplitud de información y con gran posibilidad de actualización, esto no significa, tal como menciona Cabero Almenara la generación o adquisición de conocimiento significativo, "para ello es necesario su incorporación dentro de una acción perfectiva, su estructuración y organización, la participación activa y constructiva del sujeto". [3]

La mayoría cuenta con un gran número de herramientas para el soporte de contenidos, comunicación y colaboración. Una de las plataformas más utilizadas es Moodle, la cual cuenta con un gran número de actividades y recursos disponibles. Varios estudios permiten observar que las actividades más utilizadas son los foros, wiki, consultas y cuestionarios. Le siguen en menor medida los recursos chat, glosario y diario, y con un uso casi nulo se encuentran la lección y base de datos. [4, 5, 11]

Una actividad que no se menciona en los artículos recabados es el libro, este ha sido diseñado para la creación de materiales hipermediales no muy extensos, los cuales pueden organizarse generalmente en dos niveles. Tanto docentes como estudiantes pueden además de navegarlo, imprimir una parte del mismo por ejemplo una selección, un capitulo o inclusive todo el libro. Sólo los profesores pueden crear y editar libros, por lo que la retroalimentación producto de la interacción de los estudiantes puede diseñarse fuera de él, por ejemplo a través de los foros.

1.2 PEDCO: plataforma Comahue

Desde la creación de la plataforma PEDCO [6] en el año 2004, se ha tenido como meta la promoción de la inclusión genuina de las nuevas tecnologías de la información y la comunicación con una mirada puesta en el mejoramiento de las prácticas docentes.

A través del acompañamiento desde el Sistema de Educación Abierto y a Distancia (SEADI) esto no sólo ha sido aplicado al ámbito de la Facultad de Informática sino

que además se ha extendido a toda la universidad. Los artículos escritos y presentados en distintos congresos desde esa fecha han permitido además la creación de espacios de intercambio y colaboración con docentes de otras universidades y de otros países.

Desde el inicio de la plataforma mencionada se ha tenido como objetivo el estudiar los problemas, demandas y requerimientos que las prácticas docentes instalan, y desplegar de manera integrada desarrollos tecnológicos que enriquezcan al docente y lo inciten a revisar la enseñanza. La mera introducción de tecnologías no genera en sí misma innovación sino que hace falta además investigar y promover, de ser necesario, la implementación de nuevos recursos tecnológicos.

El material utilizado en las cátedras ha evolucionado y actualmente el soporte principal ya no es el texto o apunte impreso, sino que se adicionan diversos recursos digitales, tales como los provistos por la web 2.0. Además de este cambio en el formato y soporte de los materiales usados en un curso o materia, también ha evolucionado la forma de comunicación e interacción entre los distintos participantes de un curso (docentes, alumnos, tutores, etc.) [5]

Dentro dicha plataforma las materias o cursos cuentan con un espacio propio que es administrado por los docentes del mismo. En cada uno de ellos se utilizan carpetas para organizar los materiales antes mencionados. La forma de organizarlo depende de los conocimientos, preferencias y planificación de cada cátedra, siendo posible tener carpetas o bloques separados para cada uno de los materiales educativos (carpetas para teoría, carpetas para prácticos, etc.), tal como se muestra en la figura 1.

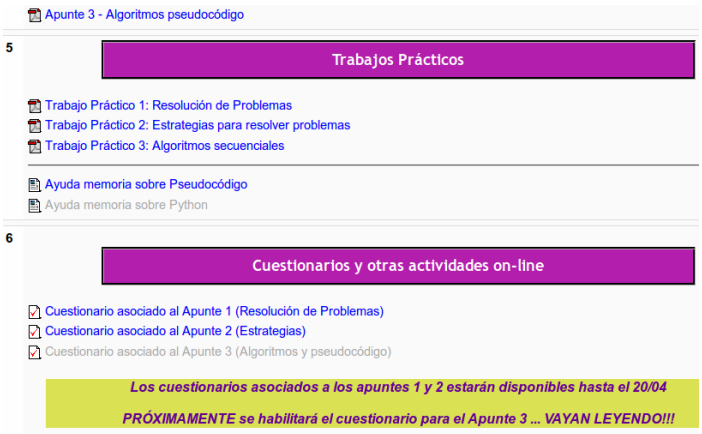


Fig.1. Organización de los contenidos por tipo de material.

También se los pueden agrupar en diferentes unidades mostrando enlaces para teoría y para prácticos dentro de un mismo bloque, y demás recursos que se deseen incorporar como se muestra en la figura 2.

Fig. 2. Organización de los contenidos por unidades temáticas.

Existen diferentes configuraciones que pueden ser utilizadas en la diagramación y estructuración de los materiales y demás elementos de un curso. La experiencia lograda desde el año 2005 permite observar que los alumnos no logran visualizar o encontrar de manera eficiente todo el material abarcado en la cátedra.

En algunos casos el uso de carpetas evitaría crear diseños sobrecargados de información, con enlaces en demasía, que en muchos casos lleva a la desorientación del alumno en cuanto a que información es nueva, de qué manera debe accederla, etc.

Tal como menciona Prensky [7] nuestros estudiantes universitarios pertenecen a una generación formada en los nuevos avances tecnológicos, en los cuales han estado inmersos desde siempre. Al estar rodeados de ordenadores, vídeos y videojuegos, música digital, telefonía móvil, otros entretenimientos y herramientas afines se han acostumbrado de manera natural a los mismos. Hoy en día la mayoría de nuestros estudiantes, y según la denominación tomada del autor mencionado, son **nativos digitales.** De todas maneras, esto no significa que puedan descuidarse las principales dificultades que tienen los estudiantes ingresantes al empezar a usar entornos virtuales: "el poder entender cómo se maneja, como así también el enviar mensajes, bajar documentos, adjuntar documentos a los mensajes y participar en foros". [8]

Dentro del conjunto de actividades disponibles en la plataforma PEDCO, se encuentra el Libro. Relevamientos sobre los usos de las actividades y diferentes recursos utilizados en todos los cursos creados en la misma han permitido observar que el uso de la actividad Libro es casi nulo. En los casos que se lo utiliza lo hacen de manera convencional.

Otro recurso no utilizado dentro de la plataforma es el blog. Los Blogs, Weblogs o bitácoras pueden definirse como "recursos informativos, en formato web, ya sea en forma textual o de imágenes, en los que una persona o grupo de personas (naturales o jurídicas), introducen por orden cronológico noticias, opiniones, sugerencias, artículos, reflexiones o cualquier otro tipo de contenido que consideran de interés, los cuales enlazan frecuentemente a otros recursos web y cuya replica está o no permitida según el propietario del Weblog ". [9]

La experiencia en su uso nos permitió pensar un uso didáctico diferente, combinándolo con el concepto de cuaderno de bitácora, logrando de esta manera

organizar y relacionar todo el material empleado en la materia en un sólo recurso. Según el Diccionario de la Lengua Española de la RAE, un cuaderno de bitácora es un: «Libro en que se apunta el rumbo, velocidad, maniobras y demás accidentes de la navegación». Nosotros hemos tomado este concepto para aplicarlo al uso que pensábamos darle al recurso libro como un lugar donde registrar los avances realizados en la materia, los temas tratados en cada clase, ejemplos mostrados, ejercicios propuestos, soluciones, discusiones surgidas en clase, dudas y preguntas pendientes, etc.

Si bien en una bitácora o blog los alumnos podrían participar a través del envío de comentarios, esta interacción en el libro se vería reflejada con el volcado de los mensajes de los alumnos producido a través de los mensajes en los foros, y la retroalimentación producto de las clases presenciales. No es una interacción directa dado que no podrían editar el libro.

2 Desarrollo de la experiencia

Se ha desarrollado una experiencia con 54 alumnos de la materia Resolución de Problemas y Algoritmos (RPA), pertenecientes al dictado del segundo cuatrimestre del año 2012. Esta materia corresponde a la carrera Licenciatura en Ciencias de la Computación, de la Facultad de Informática, Universidad Nacional del Comahue. La información recabada en la información es de tipo cualitativa.

Típicamente el material utilizado en la materia RPA se divide en presentaciones teóricas utilizadas como soporte en las clases teóricas presenciales proyectadas a través de un cañon multimedial, apuntes teóricos confeccionados por el equipo de cátedra, enlaces con referencias a material teórico adicional como libros o publicaciones disponibles en Internet, trabajos prácticos, recursos tecnológicos y enlaces a otros recursos.

Los objetivos que se han tenido en cuenta en la experiencia fueron:

Con respecto al equipo docente:

- Fomentar actividades de trabajo colaborativo y animar a la participación de todos los miembros.
- Realizar las valoraciones de las actividades realizadas.
- Facilitar y negociar compromisos de existir diferencias de desarrollo entre los miembros del equipo.
- Resolver de forma individual y colectiva las diferentes dudas que vayan surgiendo de interacción con los materiales que se le vayan presentando a los alumnos.

Con respecto a los alumnos:

- Ofrecerles un entorno más flexible para el aprendizaje, potenciando escenarios interactivos.
- Facilitar información adicional para la aclaración y profundización en conceptos.
- Ayudar a los alumnos en sus habilidades de comunicación.

- Desarrollar una evaluación continua formativa.

2.1 Organización y trabajo con la actividad libro

La actividad libro permite la creación de materiales hipermediales, los cuales pueden organizarse generalmente en dos niveles. Como se mencionó anteriormente sólo los profesores pueden crear y editar libros, por lo que para permitir la retroalimentación producto de la interacción de los estudiantes se diseñaron y enlazaron foros para consultas, comentarios, y respuestas a preguntas frecuentes.

Teniendo presente que, si bien puede estar analizado en profundidad el problema que se espera solucionar empleando el medio tecnológico, en este caso la actividad libro, y hacer uso de una de las funciones específicas de ese medio, sin embargo, no llegarse a logros educativos significativos, por no estar funcionando bien el entorno de aprendizaje que sustenta el proyecto. Entendemos por entorno de los aprendizajes el que se constituye estableciendo el rol que juega el docente, el alumno, las características del contenido que se va a construir y la integración del medio tecnológico informático, con los demás medios o recursos de aprendizaje y las interrelaciones entre todos ellos, de acuerdo con una determinada filosofía didáctica. [10]

Al comienzo del cursado de la materia se partió de un libro vacío, donde cada capítulo se empezó a construir en base al relato de lo realizado en la clase, enlazando el material utilizado en la teoría (presentaciones digitales mediante un cañon multimedial), haciendo referencias a preguntas realizadas por los alumnos sobre los conceptos abarcados en la clase, mostrando las conjeturas surgidas, conclusiones y sobre todo dificultades percibidas por el docente. Así mismo se expusieron resoluciones prácticas realizadas en forma grupal por los alumnos y se enlazaron los enunciados de los ejercicios prácticos a realizar. Todos los problemas planteados en las clases teóricas o en las prácticas se incorporaban en el libro, además de las alternativas de solución generadas en las clases presenciales y aquellas que adicionaban luego los alumnos fuera de ellas, a través de los foros.

En el libro se enlazaron:

- Relatos de lo acontecido en la clase.
- Comentarios, consultas y resoluciones a ejercicios surgidas en los foros consultas, FAQ, retención y tutorías. Estos dos últimos foros corresponden al programa de Retención estudiantil.
- Recursos digitales: apuntes teóricos-prácticos, trabajos prácticos, ejercicios resueltos, cuestionarios.
- Indicaciones sobre donde adquirir materiales impresos disponibles en centros de fotocopiado o a retirar desde la biblioteca.

El recurso libro adquiere un rol importante al posibilitar la consolidación y construcción del conocimiento compartido. Mediante la socialización los docentes equipos del equipo de cátedra, adoptan diferentes roles, manteniendo un intercambio de conocimiento sobre la base de intereses comunes y con la voluntad explícita de comunicar y compartir. Al discutir, desarrollar y contrastar sus propios puntos de

vistas, compartir sus conocimientos y tareas, elaboran un contexto discursivo que servirá de referencia para posteriores intervenciones. Esa tarea continua ha facilitado la actualización diaria de las páginas del libro, en la cual se han ofrecido materiales relevantes, nuevos recursos considerados de interés para la comunidad de la materia RPA y que al mismo tiempo, pueden ser cuestionados y reinterpretados. En la figura 3 puede verse una página del libro donde está comentado el ejercicio trabajado en clase, hay enlaces a las preguntas frecuentes, material de lectura, etc.

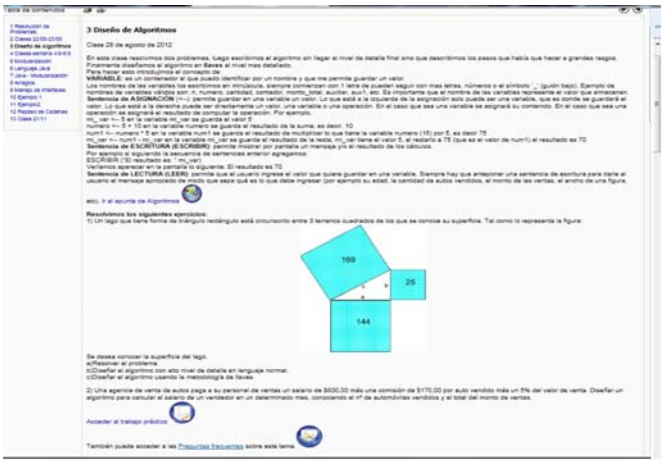


Fig.3. Página ejemplo del libro creado durante el dictado de la materia RPA.

La búsqueda de la optimización del uso didáctico del libro requirió un análisis en profundidad de todos los medios que se emplean en el proyecto educativo, tanto en su funcionalidad específica técnica, como en el uso educativo de esas funcionalidades específicas, tratando de encontrar aplicaciones nunca puestas en práctica con los otros medios. Implicó, además, la planificación de una buena articulación de los medios digitales entre sí, junto con los no digitales.

El registro de las observaciones cualitativas de todos los actores involucrados ha facilitado el tener en cuenta hallazgos imprevistos, que por lo general son factibles de suceder cuando se usa una tecnología nueva por primera vez o la misma, de forma diferente o en otro contexto.

La información cuantitativa producto de los registros realizados en la plataforma, en este caso sólo permite ver cantidad de accesos de todos los participantes a la actividad libro como a todos los recursos contenidos en la misma, además de la franja horaria en la que se accede. En esta primera experiencia se ha dado prioridad al registro de las observaciones cualitativas de todos los actores involucrados, lo que ha facilitado el tener en cuenta hallazgos imprevistos que por lo general son factibles de suceder cuando se usa una tecnología nueva por primera vez, en forma diferente o en otro contexto. El registro cualitativo se ha realizado empleando cuestionarios al final de la experiencia, asentando las observaciones de los alumnos al consultárselos en las clases presenciales y analizando los mensajes emitidos en los diferentes foros.

3 Resultados

El trabajo con el recurso libro ha permitido un intercambio fluido entre los docentes de la materia y mejorar el seguimiento de la actividad de los alumnos de la materia RPA.

Los alumnos han resaltado como muy beneficioso el contar con un único recurso que centralizaba lo desarrollado tanto en las clases teóricas como en las prácticas, enriquecido con enlaces al material de soporte utilizado en las clases, apuntes teóricos-prácticos, preguntas frecuentes, etc. Esto fue ponderado por aquellos alumnos que no han podido concurrir regularmente a las clases presenciales.

Los alumnos que cursaron la materia nos han comentado que les resultó de mucha ayuda para la preparación del final, dado que muchas veces en las clases surgen cuestiones que no están contempladas en las transparencias utilizadas, ni en los apuntes o trabajos prácticos.

Los docentes han destacado como gran ventaja de la actividad libro la posibilidad de desarrollo flexible, donde en el contenido de cada página que integran los diferentes capítulos se ha podido enlazar artículos, comentarios sobre el material referenciado, sobre lo realizado en clase, opiniones de los docentes y de los alumnos. Facilitó reflejar las ideas surgidas tanto en las clases teóricas y prácticas presenciales, como aquello gestionado fuera de los horarios de cursada.

Permitió además registrar la actividad y los avances que se iban realizando en la materia de manera en algunos casos altamente informal y extemporánea, escrita con la doble intención de provocar el diálogo y servir de bitácora como diario de aprendizaje de la asignatura (diario reflexivo de docentes y alumnos).

4 Conclusiones

En este trabajo se pretender mostrar y compartir un nuevo enfoque para el uso y trabajo con la actividad **libro** en Moodle. Desde la creación de la Plataforma PEDCO nuestra atención se ha centrado en procedimientos, estrategias, mecanismos, dispositivos y experiencias cuyo objetivo es la evaluación de los usos de las TIC para impulsar nuevas formas de aprender y enseñar.

Es común observar en la diagramación y configuración de los cursos en PEDCO el uso de páginas extensas, provistas de muchos enlaces y sobrecargada de información, que a veces al no estar bien organizada resulta dificultosa de entender hasta para los estudiantes más familiarizados con las nuevas tecnologías.

El recurso libro facilitó la organización del contenido a trabajar en la materia RPA, y adquirió un rol importante al posibilitar la consolidación y construcción del conocimiento compartido. Cada uno de los integrantes del equipo de cátedra (docentes de teoría, jefe de trabajos prácticos, ayudantes) pudo aportar sus conocimientos más relevantes, resultando un beneficio por la suma de los esfuerzos realizados.

Al finalizar el dictado como ventaja para los alumnos y docentes, el producto resultante fue un único recurso, en este caso un libro, que no sólo contenía material teórico y práctico, sino que además permitía observar lo transitado en cada clase presencial, y en algunos casos, hasta lo ocurrido fuera de ellas.

En las clases presenciales a veces se observa como una pregunta disparadora puede ocasionar que no se vean todos los temas pautados para ese día, no se alcance a mostrar todos los ejemplos planificados o surjan nuevos problemas a resolver en base a las inquietudes de los estudiantes. Los docentes al tener que escribir lo acontecido en cada clase, podían además realizar otros aportes extras producto de la reflexión que es posible cuando uno escribe un texto.

Evaluar el uso de los recursos TIC empleados en nuestra práctica docente nos permitirá identificar los procesos y prácticas que resulten ser más eficaces, además de compartir y extender el nuevo conocimiento adquirido a otros ámbitos educativos.

Referencias

1. Bustos, M. Román, La importancia de evaluar la incorporación y el uso de las tic en educación. Revista Iberoamericana de Evaluación Educativa, vol. 4, n. 2, (2011), pp. 1-5.
2. Plaza, J. La enseñanza mediada por tecnología. Primeras Jornadas de Educación Mediada por Tecnología del SEADI (2007).
3. Cabero Almenara, J. La sociedad de la información y el conocimiento, transformaciones tecnológicas y sus repercusiones en la educación. (2001)
4. Sánchez, J., Morales, S Docencia universitaria con apoyo de entornos virtuales de aprendizaje (EVA). Digital Education Review, 21, (2012) pp. 33-46.
5. Fracchia, C. Plaza, J. Análisis de los materiales educativos incorporados a la plataforma PEDCO. RUEDA. V Seminario Internacional. (2010). Universidad Nacional Del Centro De La Provincia De Buenos Aires .Tandil, Bs. As., Argentina.
6. Fracchia, C., Alonso de Armiño, A. PEDCO (Plataforma de Educación a Distancia Universidad Nacional del Comahue).X Congreso Argentino de Cs. De la Computación. Universidad Nacional de la Matanza. Vol. I. .(2004)
7. Prensky, M. Digital natives, digital immigrants. On the Horizon 9 (5). pp. 1-6. (2001)
8. Castro Chans, B., Godoy, M., Sobol, B., Mariño, S. Implementación de un EVA para el ingreso a la Universidad: El caso del módulo "Estrategias de Aprendizaje en la Universidad" para los alumnos de la Licenciatura en Sistemas de Información. VII Congreso de Tecnología en Educación y Educación en Tecnología. (2012)
9. Ferrada Cubillos, M. Weblogs o bitácoras : un recurso de colaboración en línea para los profesionales de la información. Serie Bibliotecología y Gestión de Información(6).
10. Vacca, A. Criterios para Evaluar Proyectos Educativos de Aula que incluyen al Computador. Revista Iberoamericana de Evaluación Educativa, 4(2). pp. 36-54. (2011)
11. Sánchez Santamaría, J., Sánchez Antolín,P., Ramos Pardo, F. Usos pedagógicos de Moodle en la docencia Universitaria desde la perspectiva de los estudiantes. REVISTA IBEROAMERICANA DE EDUCACIÓN. N.o 60 (2012), pp. 15-38 (1022-6508)

Uso de Software Interactivo como Facilitador para la Introducción Temprana de Conceptos de Control Robusto

Patricia Baldini[1,2], Guillermo Calandrini[1], Pedro Doñate[1], Héctor Bambill[2],

[1] Universidad Nacional del Sur, Bahía Blanca, Buenos Aires, Argentina
[2]Universidad Tecnológica Nacional, FRBB, Bahía Blanca,Buenos Aires, Argentina
{pnbaldi, calandri, pdonate, hbambill }@criba.edu.ar

Resumen. En este trabajo se presenta una propuesta pedagógica en la que el empleo de una herramienta de CAD interactiva de libre disponibilidad sumada a una experiencia de laboratorio, posibilita la introducción de nociones avanzadas de incertidumbre y robustez en un curso de control clásico para ingeniería. Estos conceptos, normalmente ajenos a una materia introductoria de sistemas de control por su complejidad matemática y limitaciones de tiempo, son incorporados intuitivamente mediante el software SISO-QFTIT que utiliza la Teoría de Realimentación Cuantitativa como marco de diseño en el dominio frecuencial. Se logra articular de modo natural los fundamentos del control clásico y robusto. Con una interfaz gráfica amigable e interactiva se adapta a las habilidades propias del sujeto educativo actual, constituyéndose en una eficaz herramienta didáctica. La evaluación realizada muestra que la herramienta informática facilita la construcción de los conceptos asociados a las distintas etapas del proceso de diseño robusto.

Keywords: CAD de control**,** simulación interactiva, educación en control, control robusto.

1 Introducción

Como parte de la formación en las carreras de Ingeniería Eléctrica y Electrónica se incluye, promediando la carrera, una asignatura básica en la temática de control de sistemas. Este curso cubre contenidos clásicos de análisis y diseño de controladores en los dominios del tiempo y de la frecuencia para sistemas lineales modelados mediante función de transferencia. Por lo general, dentro de los temas tratados se deja de lado un aspecto importante, tanto desde el punto de vista práctico como conceptual, como lo es el tratamiento de la incertidumbre que surge naturalmente en el proceso de modelado y su consecuente problemática asociada a la estabilización y el control robusto. Las razones que justifican esta omisión se encuentran en la complejidad matemática asociada al marco teórico formal y en la dificultad de hallar un balance entre el tiempo requerido para su presentación, la claridad conceptual y su aplicabilidad en el contexto y nivel de un curso inicial. De manera similar, si se analiza la bibliografía moderna de introducción a lo sistemas de control, si se contempla esta temática, se hace de un modo más bien tangencial a través de los conceptos de márgenes de estabilidad y sensibilidad paramétrica [3, 16, 17]. Como

consecuencia, el tema queda relegado para su tratamiento en cursos muy específicos, frecuentemente de carácter optativo, y en muchos casos de posgrado. El resultado concreto es que, en la construcción cognitiva de los alumnos se afianza la idea de que la obtención de un modelo único y perfectamente definido no solo es posible sino también suficiente a los efectos de su control.

En este trabajo se presenta una experiencia didáctica diseñada para introducir tempranamente los conceptos de incertidumbre y robustez, sustentada por un software CAD de uso libre, interactivo y con una interfaz gráfica amigable e intuitiva. Este software, denominado SISO-QFTIT [4] conduce al diseño de un controlador robusto basado en la Teoría de Realimentación Cuantitativa o QFT a partir de sus siglas en inglés [12,13]. Esta teoría es particularmente accesible y permite articular de manera directa y sencilla los conceptos del control clásico con los de incertidumbre de modelo y robustez proporcionando además un procedimiento de diseño transparente, versátil y práctico. Dadas las habilidades que naturalmente tienen incorporadas los alumnos en el manejo de dispositivos portátiles, la interfaz gráfica del programa es asimilada en forma inmediata no requiriendo tiempo de aprendizaje en este sentido.

La propuesta se implementa en un primer curso de control realimentado para la carrera Ingeniería Electrónica del Departamento de Ingeniería Eléctrica y de Computadoras de la Universidad Nacional del Sur. El tema se aborda de una manera completamente práctica a través de una experiencia en laboratorio donde se trabaja sobre un típico sistema de control de posición. La metodología didáctica adoptada se sustenta en el paradigma de aprendizaje basado en el descubrimiento y el trabajo colaborativo,

Los resultados de la experiencia se evaluaron mediante una encuesta realizada al final del curso, que evalúa la percepción propia de los alumnos en cuanto a los objetivos pedagógicos planteados, mostrando lo favorable de la iniciativa. Se resalta también la importancia que un software de uso libre puede tener como herramienta didáctica para facilitar la introducción temprana de conceptos relevantes en el área del control realimentado, sin necesidad en esta etapa de recurrir a complejos desarrollos teóricos . Se comprueba que este tipo de software constituye un aliado importante en la enseñanza que debería extenderse a otras áreas de la ingeniería.

En la sección 2 se resumen muy brevemente los conceptos básicos de control cuantitativo robusto. En la sección 3 se presenta el software de diseño utilizado, se detallan las características relevantes que justifican su elección comparando al mismo con otras opciones posibles tanto desde el punto de vista funcional como didáctico. En la sección 4 se describe la experiencia realizada junto con una breve descripción del sistema físico sobre el cual se realizaron las mediciones y validación de los resultados. En la sección 5 se detallan, a modo de ejemplo, un conjunto de resultados típicos obtenidos a partir del software y los resultados de la encuesta realizada y finalmente, en las secciones 6 y 7, se presentan resultados y conclusiones.

2 Fundamentos de la Teoría de Realimentación Cuantitativa

La Teoría de Realimentación Cuantitativa es una técnica práctica de diseño robusto de controladores en el dominio de la frecuencia basado en el modelo de función

transferencia. Reinterpreta las ideas de Bode del control clásico, llevándolas a una forma cuantitativa reforzando la idea de que la realimentación es necesaria en función de la existencia de incertidumbre en el modelo de la planta o por la presencia de perturbaciones no medibles actuando sobre la misma [8,9,12,13].

El objetivo de QFT es la síntesis de un controlador lo mas simple posible y con ancho de banda mínimo, en base a una función transferencia nominal. Este controlador debe garantizar que se satisfagan las especificaciones del sistema cualquiera sea la planta dentro del conjunto posible determinado por la incertidumbre, aún en presencia de posibles perturbaciones.

El esquema de control realimentado contemplado en QFT incluye dos posibles grados de libertad según el esquema de la Fig. 1. Con el controlador, $G(s)$, en el lazo cerrado se logra cumplir con las especificaciones de robustez, mientras que el pre-compensador, $F(s)$, permite ajustar la respuesta en frecuencia deseada.

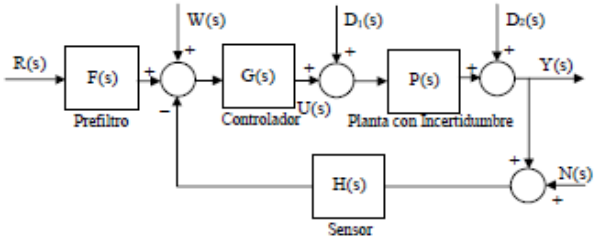


Fig 1: Esquema de control con dos grados de libertad .

Para la planta $P(j\omega)$, se define su *template* como el conjunto de respuestas en frecuencia posibles asociadas a la variación de los parámetros inciertos dentro de rangos definidos. Tanto los *templates* como las especificaciones cuantitativas de estabilidad, comportamiento temporal y rechazo de perturbaciones son trasladadas a un conjunto de curvas representadas en un diagrama polar denominado carta de Nichols. Estas curvas conocidas como contornos o *bounds*, permiten continuar con el proceso de diseño en base a la planta nominal y sirven de guía para la determinación de la función transferencia de lazo abierto nominal, $L_0(j\omega) = P_0(j\omega)\ G(j\omega)$, mediante la introducción sucesiva de ganancia, polos y ceros en el controlador $G(j\omega)$.

El controlador se sintetiza de modo de lograr que la curva de $L_0(j\omega)$ en la carta de Nichols se ajuste lo más posible a los *bounds* para cada frecuencia de interés, resultando en la minimización de la ganancia de alta frecuencia. Si $L_0(j\omega)$ satisface las restricciones, se garantiza que también lo harán todas las funciones de lazo correspondientes a las plantas del *template*.

Todas las etapas del diseño admiten su correlación gráfica de modo que su aprendizaje es ameno y con alto contenido formativo [6,7,11,14].

3 Criterio de Selección y Características del Software de Diseño

La construcción de las estructuras lógicas y formales propias de la teoría de control debe ser acompañada fuertemente por aspectos intuitivos y estrategias que forman

parte del conocimiento experto y resultan muy difíciles de poner en evidencia para que sean asimilados por los alumnos. Las herramientas de software accesibles a través de Internet resultan de gran estímulo para desmitificar conceptos matemáticos abstractos e involucrar más activamente a los alumnos en su propio proceso de aprendizaje. Actualmente, diversos paquetes de software específicos proporcionan una alternativa interesante.

Desde el punto de vista pedagógico, al considerar la incorporación de software de diseño robusto en el marco de QFT, se puede distinguir dos estrategias [6] :

-su empleo como herramienta auxiliar, donde se encara el diseño contemplando dos etapas. La primera es la síntesis o determinación de los valores del conjunto de parámetros de diseño, mientras que la segunda es el análisis o validación de resultados obtenidos en relación a las especificaciones. Esto comúnmente conduce a una nueva iteración en un procedimiento de prueba y error que puede resultar bastante tedioso.

-su utilización como herramienta interactiva, donde se combinan ambas etapas y los efectos del cambio de los parámetros del controlador son presentados en forma inmediata. En esta aproximación, el diseño se torna realmente dinámico permitiendo al usuario percibir el modo en que cada uno de los elementos que está modificando influye en el comportamiento del sistema. Esto permite orientar el diseño en la dirección de hallar un compromiso aceptable entre todos los requerimientos, generar un criterio intuitivo relacionado con el conocimiento experto, e identificar rápidamente si las especificaciones pueden o no ser satisfechas.

En general, en el campo del control automático MATLAB® representa la herramienta de software mas difundida ya que provee una gran variedad de funciones de librería o *toolbox* que implementan las diversas técnicas usadas en control, así como también, una interfaz gráficas de usuario (GUI). Existen en la actualidad diferentes herramientas de CAD orientadas a facilitar la resolución de las diferentes etapas de la metodología QFT implementadas en MATLAB [1,2,10,15]. Estas proveen un conjunto de funciones o *toolbox*, o incorporan gráficos interactivos basados en la interfaz gráfica de usuario (GUI). En el primer caso, el más versátil, los usuarios deben tener algún conocimiento básico del lenguaje de programación específico ya que es necesario reescribir líneas de código para resolver cada problema particular. En el segundo caso, mas estructurado, las funciones se ejecutan directamente, ingresando los datos requeridos en los campos disponibles para comenzar el cómputo. Dentro de las herramientas mas difundidas y completas se mencionan los *toolboxs QFT Frequency Domain Control* (FDCDT) de tipo comercial [2] y *QFT Control* (QFTCT) de libre acceso [10] . De todos modos, en la primera el grado de interactividad es algo limitado y, en ambos caso, se requiere de la instalación de MATLAB, lo que implica contar con una licencia muy costosa.

Por otra parte, en el campo del software de distribución gratuita se encuentra disponible SISO-QFTIT [4,5] (Single Input Single Output Quantitative Feedback Theory Interactive Tool) caracterizada por su alta interactividad en cada etapa del proceso y su facilidad de uso [5]. La operación directa con el puntero del mouse sobre los diferentes elementos presentes en la ventana de aplicación permite interconectar visualmente sus consecuencias.

Desarrollada en ambiente *Sysquake* [19], se presenta en forma de archivo ejecutable bajo sistema operativo Windows y Mac proporcionando alta portabilidad.

Su uso facilita tanto la comprensión de los conceptos básicos involucrados en la metodología tratada como el desarrollo de las habilidades fundamentales de diseño y el sustento para la base teórica necesaria. Los efectos de cada acción del usuario introducida por pantalla durante el proceso de diseño, se reflejan en cambios inmediatos de todos las figuras de la ventana gráfica, respondiendo al concepto de *dynamics pictures* [14,20].

Una limitación de este software es que está restringido a sistemas con una entrada y una salida y en el tipo de parámetros inciertos admisibles. También puede mencionarse que la calidad de los gráficos es menor que la de otro casos. De todos modos, se ajusta perfectamente a los objetivos planteados.

4 Metodología Implementada

Se plantea el problema de diseñar un sistema de control de posición a partir de un motor de corriente continua de imanes permanentes (Fig. 2)[8]. Con este objetivo cada grupo de estudiantes seleccionan los componentes disponibles; reconocen las leyes físicas implicadas para determinar la estructura del modelo y diseñan los experimentos necesarios para la identificación de sus parámetros. La dispersión natural en los resultados permite introducir el concepto de incertidumbre estructurada y la dispersión dentro del espacio paramétrico se establece en base al intercambio de resultado entre grupos de trabajo. Queda en claro el hecho práctico de que el modelo es solo una aproximación del sistema físico de modo que no es posible determinar valores exactos.

Las limitaciones del sistema adoptado como tensiones máximas admisibles en la alimentación del motor, torque máximo, etc., conducen a determinar el conjunto de especificaciones de funcionamiento. La estructura del controlador adoptada es la clásica del PID que permite la comparación con los resultados de técnicas convencionales de diseño. En este punto, se plantea el cuestionamiento sobre la influencia de la incertidumbre del modelo en los resultados esperados y la posibilidad de incluirla en el proceso de diseño. Como una alternativa válida en esa dirección, se presenta la metodología enmarcada en QFT que conduce a un controlador robusto.

El modelo adoptado es de segundo orden con una función transferencia dependiente de dos parámetros: la ganancia y un polo real ($K \approx [135, 224]$, $p \approx [4.5, 8]$).
Como guía en la resolución del problema se platean la siguientes tareas:

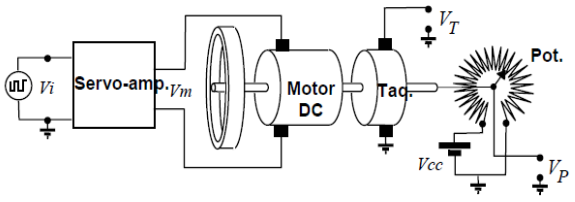


Fig.2: Esquema del sistema de control de posición a lazo abierto

- Definir el modelo de incertidumbre paramétrica.

- Determinar el rango de frecuencias de interés.
- Determinar la planta nominal y generar los *templates* con SISO-QFTIT
- Traducir las especificaciones del dominio tiempo al de la frecuencia.
- Analizar la necesidad de restricciones para tener en cuenta estabilidad, esfuerzo de control, rechazo de perturbación y considerar restricciones en la señal de control para evitar saturación del servoamplificador y la fricción estática en el motor.

-Generar los *bounds* y su intersección (usar herramienta de CAD SISO-QFTIT)
-Sintetizar el controlador PID del lazo (loop-shaping) con el software interactivo.
-Simular los resultados
-Considerar la opción de ajustar las restricciones.
-Validar el diseño en el laboratorio.

5 Ejemplo de Posibles Resultados Gráficos

La metodología QFT está implementada en distintas etapas:

Etapa 1: Ingreso del modelo y generación de los *templates*
En la primera ventana, el usuario define la planta nominal mediante la ubicación gráfica de polos y ceros, la incertidumbre de sus elementos y frecuencias de trabajo. Inmediatamente se muestran los *templates.* (Fig. 3)

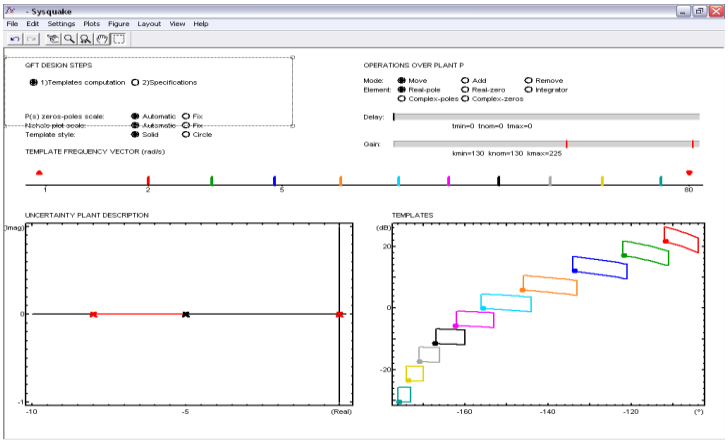


Fig. 3: Ventana de Inicio: definición del modelo y generación de *templates*

Etapa 2: Especificaciones en frecuencia.
Se selecciona y configura el tipo de especificaciones diseño entre seis opciones posibles junto al rango de frecuencias en que se debe cumplir cada una. En forma automática se generan los *bounds* asociados, pudiendo ser visualizados en forma individual, conjunta o intersectados, según se observa en la Fig. 4.

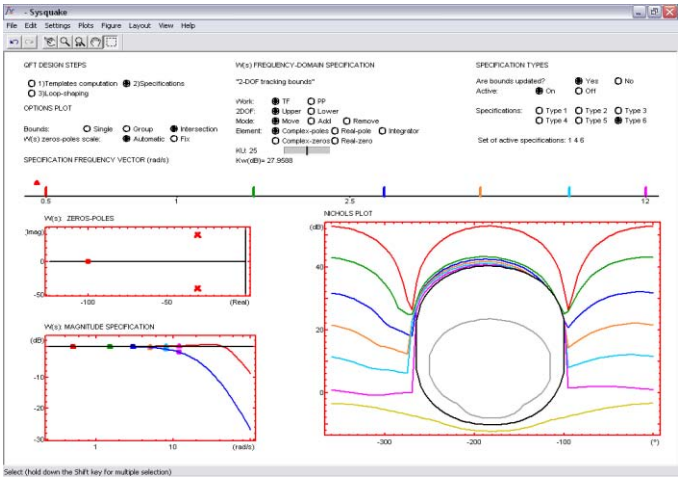


Fig. 4. Ventana: generación e intersección de *bounds*

Etapa 3: Síntesis del controlador o *Loop shaping*
En esta ventana (Fig. 5) se realiza la síntesis del controlador en base a su función transferencia y sobre la carta de Nichols. Se ingresan gráficamente los elementos componentes (ganancia, polos y ceros) con la inmediata visualización de su efecto sobre la respuesta en frecuencia de lazo abierto. En este caso se utlilizo una estructura PID convencional. De ser necesario, el ajuste final puede lograrse con un prefiltro en una etapa opcional.

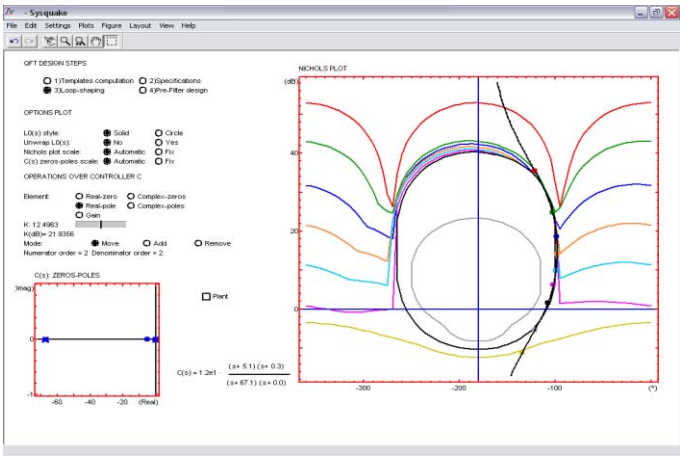


Fig. 5. Ventana: diseño del controlador sobre el gráfico de Nichols

Etapa 4: Validación del diseño
La ventana final muestra los gráficos logrados al incluir el controlador en el lazo y el posible prefiltro. Se pone en evidencia el grado de cumplimiento de las

especificaciones con las curvas representativas en tiempo y frecuencia. En cada gráfico se incluyen los peores casos resultantes teniendo en cuenta la incertidumbre paramétrica y los límites fijados por las especificaciones (Fig. 6).

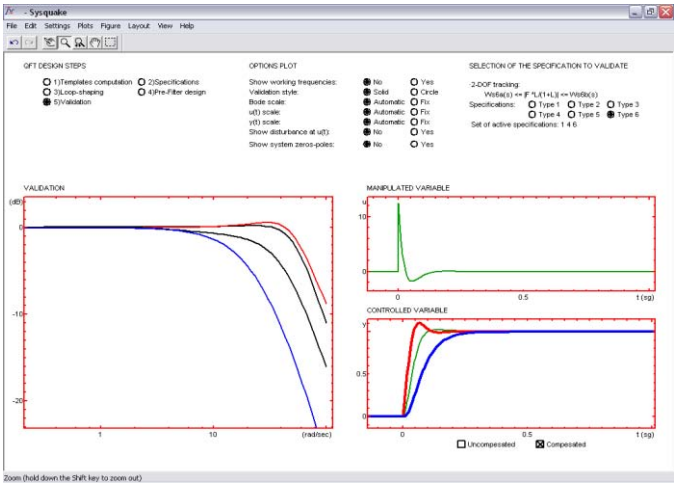


Fig. 6: Verificación de especificaciones con el controlador diseñado y el posible pre-filtro

6 .Resultados Obtenidos

La experiencia se desarrolló con diez alumnos que optaron por esta metodología integrando tres grupos de trabajo. Los resultados obtenidos se evaluaron desde la óptica de los alumnos mediante una encuesta de percepción realizada al finalizar la experiencia, con los siguientes resultados:

- 100% detecta la incertidumbre y la dispersión en los resultados de los valores experimentales.
- 50% considera que esta incertidumbre puede afectar al rendimiento del sistema y debe ser tenido en cuenta en el diseño.
- 30% considera que el desajuste entre el modelo y el sistema es una consecuencia de los efectos no lineales y otros errores de modelado y no necesariamente debido al ruido asociado a las medidas.
- 70% considera que resulta significativo tener en cuenta la incertidumbre en el proceso de diseño.
-100% considera que el paquete de software QFT es intuitivo y fácil de usar.

Por otra parte, desde el punto de vista formativo y según la la opinión de los docentes, se reconocen los siguientes avances:

- Mediante el paradigma del aprendizaje por descubrimiento, los alumnos reconocen la incertidumbre propia de todo proceso de modelado y la asocian en este caso a una dispersión en la determinación de los parámetros del modelo.

- Se introduce naturalmente una metodología robusta, conceptualmente sencilla y básicamente gráfica, que amplía la visión de los alumnos en cuanto a las limitaciones propias del control clásico.
- Se orienta el aprendizaje hacia un conocimiento experto.
- Los alumnos observan que el diseño robusto, no necesariamente óptimo, se comporta bien cumpliendo las especificaciones de comportamiento requeridas sobre toda la gama de posibles variaciones, siempre que esas especificaciones sean razonables teniendo en cuenta los componentes del sistema.
- Motivados por una "necesidad de conocer" los estudiantes atribuyen valor y significado a su proceso de aprendizaje, se comprometen en las diferentes tareas así como en la comprensión de las ideas que las sustentan.

7 Conclusiones

En este trabajo se describió una experiencia didáctica en la que se conjugan una experiencia de laboratorio y una herramienta de CAD interactivo como facilitador para la introducción, en un curso de control clásico, de modelado con incertidumbre y el consecuente diseño robusto.

La metodología presentada permite una mayor independencia y compromiso del alumno con su propio proceso de aprendizaje y está centrada en las competencias que debe proveer la asignatura de control. Se favorece el desarrollo del aprendizaje autónomo, fortaleciendo la habilidad para evaluar herramientas de diseño, apreciar las limitaciones de un modelo, incorporar los conceptos de incertidumbre y *loop-shaping*.

La teoría del control clásico sirve de base suficiente para la comprensión de QFT, haciendo viable su incorporación en tiempos razonable.

En las diferentes etapas del proceso el nivel de interactividad de SISO-QFTIT garantiza un mayor y más rápido nivel de entendimiento de la temática de control robusto, combinando las fases de análisis y diseño. Se pone en evidencia mediante la percepción visual en qué dirección variar los parámetros para lograr los objetivos y la correspondencia entre los dominios de la frecuencia y el tiempo .

El uso de un sistema sencillo y la herramienta de CAD adoptada conforman una propuesta didáctica motivadora que mejora la rápida comprensión de las etapas del diseño y de los conceptos que lo sustentan. La experiencia obtenida sobre sistemas simples puede ser fácilmente generalizada a situaciones más complejas, y permite un conocimiento intuitivo de los formalismos matemáticos subyacentes.

Por otro lado, se crea conciencia en relación a la importancia de la existencia de software de libre distribución y su incorporación a la enseñanza, teniendo en cuenta el alto costo de los comerciales de uso frecuente en control, muchas veces inaccesible para la Universidad pública.

References

1. Barreras, M., Vital, P. y García-Sanz, M.: Interactive tool for easy robust control design. Prooc. of the IFAC Internet Based Control Education , Madrid (España),pp: 83-88 (2001)

2. Borghesani,C., Chait,Y. and Yaniv,O. : Quantitative Feedback Theory Toolbox for use with MATLAB. The MathWorks Inc, Natick, MA, (1995).
3. Dorf, R.C., Bishop, R.H.: Modern Control Systems, Chapter 12, 12th Ed., Prentice Hall, New York. (2010)
4. Díaz, J.M., Dormido, S., Aranda, J. http://ctb.dia.uned.es/asig/qftit/principal.html, UNED (2004)
5. Díaz, J.M., Dormido, S. y Aranda, J.: An Interactive Software tool to learn robust control design using the QFT methodology. (2007)
6. Dormido, S .; The role of interactivity in control learning, Int. Jou. Engin. Ed. 21(6) pp. 1122-1133 (2005)
7. Dormido, S., Gordillo, F., Dormido Canto, S. y Arancil, J.: An interactive tool for introductory nonlinear control systems education. 15th IFAC World Congress. Barcelona, España. (2002)
8. García Sanz, M.: Quantitative Robust Control Engineering: Theory and Applications. Educational Notes RTO-En-SCI-166, pp.1-44 (2006)
9. García Sanz, M.: Control Robusto Cuantitativo QFT: historia de una idea. RIAI.2 -3. pp. 25- 38 (2005)
10. Garcia-Sanz, M., Mauch, A. and Philippe, Ch.: The QFT Control Toolbox (QFTCT) for Matlab, CWRU, UPNA and ESA-ESTEC, Version 3.31, November (2012) http://cesc.case.edu/OurQFTCT.htm
11. Guzman, J.L., Costa Castelló, R., Dormido, S. y Berenguel, M.: Study of fundamental control concepts through interactive learning objects. 18th IFAC World Congress. Milán. Italia pp. 7286-7291 (2011)
12. Horowitz, I.M.: Quantitative Feedback Design Theory-QFT,QFT Publishers, Denver (1993)
13. Houpis, H., Rasmussen, S.J. y García Sanz, M.: Quantitative Feedback Theory: Fundamentals and Applications, 2da. Ed, CRC Press, Florida (2006)
14. Johansson M., Gäfvert, M. y Ämtröm, K.J.: Interactive Tools for Education in Automatic Control. IEEE Control Systems Magazine. 18(3), pp. 33-40 (1998)
15. Nandakumar , R.and Halikias, G. D.: A new educational software tool for robust control design using the QFT method. Proceedings of the 42nd IEEE Conference on Decision and Control . Maui, (Hawai USA),pp. 803 -808 (2003)
16. Ogata, K. : Ingeniería de Control Moderna, Cap. 10, 4ta. Ed. , Prentice Hall, Madrid.(2011)
17. Paraskevopoulos, P.N.: Modern Control Engineering, Ch 15, Marcel Dekker. New York.(2002)
18.Perrenet, J.C., Bouhuijs, P.A.J. & Smits, J.G.M.M.: The suitability of problem based learning for engineering education: theory and practice. Teaching in higher education, 5, 3, pp. 345-358. (2000)
19. Piguet, Y. : SysQuake: User's manual. Calerga. Lausanne Federal Polytechnic School Automatics Institute (1999) www.calerga.com/products/Sysquake
20. Wittenmark, B. Häglund, H. y Johansson M.: Dynamic pictures and interactive learning, IEEE Control Systems Magazine. 18(3), pp. 26-32 (1998)

TICs para una Educación Inclusiva

Emilce Castillo[1], Rossana Sosa Zitto[1], Ulises Rapallini[1], Rafael Blanc[2], Leandro Lepratte[2]

[1] Universidad Autónoma de Entre Ríos, Facultad de Ciencia y Tecnología.
[2] Universidad tecnológica Nacional, Facultad Regional Concepción del Uruguay.

Resumen: El presente trabajo permite desarrollar el Núcleo de Investigación en Tecnologías para la Inclusión Social de la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos en la Sede Concepción del Uruguay. El mismo tiene previsto impulsar un proyecto sobre Modelos interactivos de aprendizajes basado en tecnologías para la inclusión social de bajo costo y aplicaciones múlti-usuarios. Cuenta con el apoyo del IproDi, Instituto Provincial de Discapacidad de la Provincia de Entre Ríos y con la Dirección de Discapacidad de la Municipalidad de la ciudad de Concepción del Uruguay, Provincia de Entre Ríos, esto asegura que los resultados del mismo sean integrados en una línea estratégica de desarrollo de tecnologías para la inclusión social de personas con discapacidad, y refuerza también la posibilidad de sostener al entramado de instituciones que se vinculan al proyecto ampliándose a partir de las actividades de transferencia y difusión. La Universidad Autónoma de Entre Ríos cuenta con experiencia en el área discapacidad y en el área de desarrollo social desde hace varios años, en el área inclusión social ha desarrollado los proyectos "Juego para todos" y "Motorización de Sillas de Ruedas Convencionales", los cuales fueron presentados en Tecnópolis en el marco de los festejos del Bicentenario.

Palabras claves: tics, educación inclusiva, discapacidad, tecnología asistiva.

1. Introducción

En Argentina desde hace años se ha emprendido un camino ambicioso el de sentar las bases para una educación pública inclusiva y de calidad, hacer una escuela que desafíe las diferencias, que profundice los vínculos y que permita alcanzar mayor igualdad social y educativa para los jóvenes. En este marco las instituciones de la Provincia de Entre Ríos orientadas a la educación y promoción de la salud de niños y jóvenes discapacitados han evidenciado crecientes esfuerzos por integrar las TIC a las actividades pedagógico - terapéuticas que llevan adelante. En especial estas iniciativas se han centrado en la facilitación de la infraestructura para la accesibilidad. Estas acciones también se evidencian en instituciones privadas y ONGs. Del estado de situación analizado, se han encontrado aisladas experiencias virtuosas de aplicación de TIC facilitadoras de procesos de comunicación en niños con discapacidad en la provincia de Entre Ríos. La intención clave de este proyecto es iniciar un desarrollo regional endógeno de estas tecnologías, articulando instituciones de CyT, gubernamentales y privadas de apoyo a personas con discapacidad, para incorporar estas tecnologías al mayor número posible de instituciones de la provincia. En respuesta a esta iniciativa, demandada socialmente por diversas instituciones, articuladas ellas por el Instituto Provincial de la discapacidad (IproDi), se decide impulsar desde la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos, Sede Concepción del Uruguay un equipo de I+D para el diseño e implementación local de este tipo de TICS. En el segundo apartado se hablará sobre educación e inclusión, en el siguiente se tratará la situación de la provincia de Entre Ríos en aspectos de discapacidad, en la cuarta sección se situará al lector en la relación entre las TICS y la mejora que puede presentar para las personas con discapacidad. En el quinto apartado se plantea el proyecto con el cual se piensa mejorar la situación de las personas con discapacidad, y finalmente se aportará unas palabras finales a modo conclusión de este ensayo.

2. La Educación Inclusiva

La educación inclusiva responde a un enfoque filosófico, social, económico, cultural, político y pedagógico que persigue la aceptación y valoración de las diferencias en la escuela para cada uno de los alumnos. En la educación inclusiva los alumnos se benefician de una enseñanza adaptada a sus necesidades. Dentro de este marco se plantea la necesidad de repensar la práctica docente, proponiendo nuevos desafíos que permitan generar, entre otros aspectos, estrategias pedagógicas alternativas para la construcción de: respuestas a las necesidades educativas para las personas con barreras para el aprendizaje, su participación en distintos contextos, la promoción de las alfabetizaciones múltiples y el aprendizaje constructivo. (Zappalá, Köppel, Suchodolski, Octubre 2011).

La educación especial es la modalidad del sistema educativo destinada a asegurar el derecho a la educación de las personas con discapacidades, temporales o permanentes, en todos los niveles. En este contexto, el desarrollo de proyectos que incorporen la

utilización de tecnologías de la información y la comunicación (tic) puede facilitar una mejora cualitativa de los procesos de enseñanza y de aprendizaje, desarrollar capacidades y competencias, atender a la singularidad y a las necesidades individuales de cada alumno y potenciar motivaciones que den un carácter significativo a los aprendizajes. (Zappalá, Köppel, Suchodolski, Septiembre 2011).

3. Situación actual en la Provincia de Entre Ríos

La provincia de Entre Ríos cuenta con 1.236.300 habitantes según datos del último censo Nacional de Población Hogares y Viviendas del año 2010 (INDEC), de los cuales el 11% de los habitantes, son personas con algún tipo de discapacidad (Departamento de Sistemas de Información del SNR en base al Registro Nacional de Personas con Discapacidad).
Según la estadística correspondiente al año 2011, es posible observar que el 78% de la población con discapacidad en la provincia se concentra en el tipo de discapacidad mental y motora, siendo 46% discapacitados intelectuales, y el 32% corresponde a aquellos que poseen una discapacidad de tipo motora. El resto se divide entre discapacidades auditiva y visual. Una importante proporción de estas personas son niños con alguna discapacidad que requieren de una permanente acción educativa y de promoción social que les permita concretizar sus derechos humanos fundamentales. Esto implica, entre otras cuestiones, resolver los problemas de las diversas barreras sociales, culturales, comunicacionales y artefactuales con las que tienen que interactuar y que les impiden su participación plena y efectiva en la sociedad, en igualdad de condiciones con las demás.
En particular en este proyecto se abordará la problemática comunicacional de niños, jóvenes y adultos con discapacidad producida por encefalopatías (parálisis cerebral, distrofias), síndromes de causas genéticas (Síndrome de Down, Angelmann, Weist, X Frágil), trastornos generalizados del desarrollo (autistas, Síndrome de Rett, Síndrome de Asperger), sordos e hipoacúsicos y retrasos mentales. La comunicación es una cuestión central para resolver las barreras antes mencionadas. Y tal como lo expresa la Convención sobre los derechos de personas con discapacidad la comunicación incluye: los lenguajes, la visualización de textos, el Braille, la comunicación táctil, los macrotipos, los dispositivos multimedia de fácil acceso, así como el lenguaje escrito, los sistemas auditivos, el lenguaje sencillo, los medios de voz digitalizada y otros modos, medios y formatos aumentativos o alternativos de comunicación, incluida la tecnología de la información y las comunicaciones de fácil acceso.

4. Tics y Discapacidad

Se parte de enmarcar el proyecto en las denominadas Tecnologías Sociales que son aquellas que "comprenden productos, técnicas o metodologías replicables, desarrolladas en la interacción con la comunidad y que representan efectivas soluciones de transformación social" (Dagnino, 2010).
Se tiene en cuenta el concepto de Tecnología Asistiva que define como: “tal a todo elemento de asistencia, parte de un equipamiento o sistema de productos, adquirido

comercialmente, modificado o hecho a medida, que es utilizado para aumentar, mantener o desarrollar las capacidades funcionales del individuo con discapacidad" (The Technology – Related Assitance for Individuals with Disabilities. Act of 1988).
Por último se suma el concepto de Comunicación Aumentativa Alternativa como conjunto de herramientas, estrategias y símbolos que favorecen la comunicación en personas que no disponen de un habla funcional.
Este proyecto posibilita la inclusión socio-técnica y la democratización de los procesos de co-construcción de tecnologías que satisfacen la necesidad de la población con discapacidad con déficit en la comunicación. La incorporación de las TIC de fácil acceso y los medios y formatos aumentativos y alternativos de comunicación como facilitadores de la construcción de vínculos interpersonales de las personas con discapacidad con el medio donde desempeñan actividades de la vida diaria, lleva al cumplimiento de un derecho fundamental que es el logro de una progresiva autonomía, elevar su autoestima, mejorar la calidad de vida, aspectos fundamentales para una verdadera inclusión social. Problema al cual también se busca contribuir a solucionar con la propuesta de este proyecto. El uso de los servicios, equipos o adaptaciones de Tecnología Asistiva permiten que, desde edades tempranas, las personas con discapacidad aprendan a conocer el entorno, logren mayor y mejor acercamiento a actividades de interés, como la participación en las reuniones familiares, escolares, sociales, entre otras; también, que se les dé un acercamiento más allá de su habitación o el cubículo terapéutico, con lo que se logra la interacción con los demás, para ampliar las experiencias del individuo, y, con ello, reforzar su autoestima y calidad de vida en general.
Las personas con discapacidad no necesariamente tienen que vivir una vida de aislamiento e incomunicación, según el grado de la misma muchos de ellos ya tienen la posibilidad de relacionarse con los otros. Desde hace muchos años se ha descubierto que la capacidad de los pacientes para comunicarse existe, con modos alternativos. Con entrenamiento y con ayuda de la Tecnología Asistiva diseñada a estos efectos, las personas con necesidades complejas de comunicación pueden aprender a dialogar con sus padres y sus terapeutas.
No obstante, las TIC en los términos de usos antes mencionados, han sido incorporadas a la lógica del mercado y por lo tanto en los términos de consumo, usos y valores que este establece. El elevado costo de recursos de alta tecnología para discapacidad genera un problema de accesibilidad por parte de numerosas familias con personas con discapacidad conforme a los datos antes expuestos. En esta cuestión el rol de las instituciones públicas y ONGs juegan un papel fundamental como promotores y aseguradores de los derechos de las personas con discapacidad.
La integración interinstitucional lograda para la formulación de este proyecto pretende, bajo un proceso de adecuación sociotécnica, iniciar un sendero de desarrollos de Tecnologías Asistivas para la inclusión social.

5. Proyecto

5.1. Alcance del Proyecto

En particular el alcance de este proyecto será el de diseñar, poner a prueba en interacción permanente con sus usuarios directos e indirectos y transferir bajo una

lógica de retroalimentación evaluativa, un prototipo de comunicador pictográfico de alta tecnología y un hardware adaptado para personas con discapacidad motriz; con aplicación inalámbrica. El cual será acompañado por un manual de procedimiento y prácticas educativas y comunicacionales del mismo. Junto a un programa de transferencia tecnológica (no lineal) para ser incorporado progresivamente en diferentes contextos institucionales de la provincia, en articulación con el Instituto Provincial de la Discapacidad.
El modelo de trabajo a desarrollar implica la inclusión de las familias y personas con discapacidad, e instituciones educativo-terapéuticas, promotores de salud y educadores relacionados a ellos, desde el momento de inicio del diseño del prototipo y las prácticas implícitas al mismo que se plasmarán luego para su transferencia y replicabilidad. De esta forma se establece una lógica interactiva en la co-construcción de la tecnología, democratizando el acceso a la misma desde su formulación y no como mero producto "enlatado".

5.2. Justificación del Proyecto

Las instituciones provinciales orientadas a la educación y promoción de la salud de niños y jóvenes discapacitados han evidenciado crecientes esfuerzos por integrar las TIC a las actividades pedagógico - terapéuticas que llevan adelante. En especial estas iniciativas se han centrado en la facilitación de la infraestructura para la accesibilidad. Estas acciones también se evidencian en instituciones privadas y ONGs. Del estado de situación analizado, se han encontrado aisladas experiencias virtuosas de aplicación de TIC facilitadoras de procesos de comunicación en niños con discapacidad en la provincia. De ahí que la intención clave de este proyecto sea, iniciar un desarrollo regional endógeno de estas tecnologías, articulando instituciones de CyT, gubernamentales y privadas de apoyo a personas con discapacidad, para incorporar estas tecnologías al mayor número posible de instituciones de la provincia. En respuesta a esta iniciativa, demandada socialmente por diversas instituciones, articuladas ellas por el Instituto Provincial de la discapacidad (IproDi), se decide impulsar desde la Facultad de Ciencia y Tecnología (sede Concepción del Uruguay) un equipo de I+D para el diseño e implementación local de este tipo de tecnologías. Esta iniciativa responde a una línea estratégica de la política de CyT de la Universidad Autónoma de Entre Ríos que corresponde a la permanente generación de iniciativas orientadas a la resolución de problemas sociales de la provincia. La conformación del mismo, se hace desde la perspectiva de las tecnologías sociales antes mencionadas, impulsando la inclusión socio-técnica y la democratización de los procesos de co-construcción de tecnologías. De ahí que exista a la base de la conformación del equipo una heterogeneidad de actores involucrados, de tipo institucional como así también disciplinar. Desde el punto de vista institucional participan miembros de la comunidad académica y científica, de instituciones orientadas a la discapacidad (públicas y privadas) como así también decisores políticos. En cuanto a lo disciplinar participan: ingenieros especialistas en TIC y comunicación, referentes y especialistas en discapacidad, sociólogo orientado a estudios sociales de la CyT, pedagogos, promotores de salud, especialista en formulación de proyectos de inversión, entre otros.

5.3. Etapas del Proyecto

El proyecto se desarrollará en tres grandes etapas interrelacionadas: diseño del prototipo de comunicador pictográfico y hardware inalámbrico, elaboración de la Guía uso del comunicador y de la Guía prácticas de usos comunicaciones y educativos, y finalmente de la transferencia sustentable e interactiva. La etapa de desarrollo del prototipo de Comunicador Pictográfico incluirá: investigación de información técnica sobre tecnologías aumentativas y alternativas para comunicación de personas con discapacidad, en especial las que trabajan con Sistemas Pictográficos de comunicación (SPC), aquellos orientados a facilitar la comunicación en sujetos no orales con dificultades motrices y auditivas. Luego se establecerá la definición de requerimientos del sistema, diseño técnico, programación y prueba, instalación y proceso de evaluación y adaptaciones progresivas conforme diversidad de requerimientos. Se trabajará en un dispositivo hardware inalámbrico que consiste en un pulsador inalámbrico. Y un software para el comunicador pictográfico elaborado. La característica más importante del pulsador es que tiene una comunicación inalámbrica; es por esto que consta de dos partes; por un lado estará el pulsador (que será quien transmita cuando haya señal) y el receptor conectado mediante USB a la computadora. El pulsador será realizado en plástico resistente. La parte en donde el usuario oprima será de un material de goma para que su textura realce el tacto del usuario; además contará con una luz que se encenderá cuando se transmite el pulso con el que se controlará el software; además de ser un efecto llamativo al usuario. La base del mismo se realizará con un material antideslizante, para evitar caídas o golpes accidentales. El motivo del diseño de materiales responde a la necesidad de realizar un dispositivo de poco peso y resistente. Como se detalló anteriormente el pulsador estará divido en dos partes; por un lado estará el pulsador; que constará de un microcontrolador HC908QB8; el cual realizará el trabajo de codificar la señal digital que se generará en el potenciómetro que se activa al presionar el pulsador; en una señal analógica. Dicha señal de formato binario, será modulada y enviada por el Transmisor TWS-BS-3(433) 433.9MHZ; mediante radiofrecuencia al receptor que estará conectado al puerto usb de la computadora a utilizar.

Figura 1: pulsador del dispositivo.

Esta señal de radiofrecuencia modulada llegará al receptor RWS-434N-6 433.9MHZ 5.7mA -116dB y será enviada a un microcontrolador S08JM16CLC el cual descifrará el código recibido y convertirá esta señal serial en USB, así de esta manera ingresará por dicho puerto a la computadora y manejará el software. Ambos Microcontroladores serán programados con Freescale Codewarrior; contando con varias ventajas de programación, entre las que se destacan que se podrá controlar la sensibilidad

necesaria para cada pulsador, dependiendo de las necesidades del usuario, y que se podrá configurar un código único para cada pulsador. De esta manera se obtendrá la ventaja de poder trabajar con varios pulsadores a la vez en una misma área de trabajo sin interferencias de señales. La energía con la que funcionará el pulsador estará dada por una batería de 9 Volts; utilizando los reguladores de voltaje correspondientes para el funcionamiento del micro y del transmisor.
El comunicador pictográfico estará desarrollado en lenguaje de alto nivel HTML5, así de esta manera el programa podrá ser usado directamente desde internet, contando con la opción si se desea de descargarlo e instalarlo en una computadora.
Su función principal será la de traducir imágenes en texto, formar frases y luego tener la opción de ser reproducidas en audio. Así de esta manera el interlocutor no necesita estar mirando continuamente la pantalla, sino que puede escuchar lo que el usuario quiere expresar.

Figura 2: ejemplo de interfaz prevista.

Además de la función de traductor, tendrá la opción de "chat" para poder comunicarse con otras personas que se encuentren en otro lugar. También constará de aplicaciones didácticas en donde se podrá entretener con juegos, cuentos y demás actividades que ayuden a su desarrollo intelectual y cognitivo.
Para poder ser utilizado por el pulsador adaptado, las imágenes tendrán un barrido automático, es decir irán pasando automáticamente en un intervalo de tiempo, y cuando el usuario quiera elegir una simplemente presionará el pulsador, luego de elegir la imagen y que esta se transforme en texto, volverán a pasar las imágenes para seguir eligiéndolas. Para una encontrar fácilmente los pictogramas, estarán agrupados por categoría.

5.4. Implementación

Las instituciones donde se efectuarán las actividades piloto de pruebas y evaluación serán de la ciudad de Concepción del Uruguay y Colón. Siendo beneficiarios directos de estos los niños y jóvenes pertenecientes a 2 instituciones públicas y 2 ONGs. Instituciones donde se desarrollarán las actividades de transferencia y capacitación.
En interacción con la etapa de diseño del software se trabajará, en la etapa 2 de elaboración de las Guías de uso del comunicador, y las Guías de prácticas de usos comunicacionales y educativos del mismo.
La etapa de transferencia sustentable e interactiva: corresponde a las actividades de replicabilidad y difusión de los conocimientos y prácticas generadas, como así también a la búsqueda y definición de líneas de financiamiento y acciones de política

estatal para el impulso de esta iniciativa a una escala provincial. La viabilidad de esta etapa se encuentra fuertemente apuntalada por la participación activa en el proyecto de Instituto Provincial de la Discapacidad quien considera a la inclusión de TIC como línea estratégica de desarrollo de sus políticas educativas y de promoción de la inclusión para sus beneficiarios directos e indirectos.
Se desarrollarán 3 Talleres de Transferencia y Formación en TIC para la inclusión social y autonomía. Las mismas se desarrollarán en las ciudades de Concepción del Uruguay (costa este de la provincia), Villaguay (centro de la provincia) y Paraná (Costa oeste de la provincia). Como complemento final de esta etapa se elaborará un programa de sustentabilidad para asegurar la transferencia y utilización de esta tecnología en otras instituciones de la provincia. Esta etapa parte del supuesto fundamental de que se espera lograr con el desarrollo de esta tecnología, un proceso de start up, para promover una línea de financiamiento de mediano y largo paso por parte del Estado provincial para la producción de estos comunicadores a una mayor escala y su transferencia a instituciones públicas inicialmente y ONGs de la provincia orientadas a actividad con niños y jóvenes con discapacidad.

6. Conclusiones

Es innegable el impacto de las TIC en la vida diaria de personas, industria y la comunidad en general. Actualmente estas tecnologías también están sirviendo como herramienta de integración para las personas con discapacidad demostrando su contribución al mejoramiento de su calidad de vida al facilitarles su interacción con el mundo y representan un factor que contribuye a la equiparación de oportunidades de dichas personas.

Si bien a nivel internacional el desarrollo de aplicaciones tics para personas con discapacidad se encuentra muy extendido y tiene mucha penetración dentro de los países desarrollados, no es así en países en desarrollo como la República Argentina. Esto se debe en parte a que las tecnologías actuales suelen ser muy costosas, lo que impide que los interesados puedan adquirirlas debido a que los centros de producción se encuentran en países europeos y en Norteamérica. La gran mayoría de estas tecnologías son poco conocidas en nuestro país, sin embargo proyectos como estos abren la posibilidad de desarrollar este tipo de tecnologías en el ámbito local.

De ser positivo el desarrollo del proyecto podrá ser a futuro una nueva industria que supla una carencia social como es el uso de tecnologías que permitan a los discapacitados estar más incluidos en el tejido social. Por otra parte permitirá que los especialistas en Tics tengan un segmento laboral nuevo en el cual aplicar sus conocimientos y generar beneficios para ellos y su sociedad.

7. Referencias

Bryant, Brian R. y Seay Penny Crews. The Technology – Related Assitance for Individuals with Disabilities. Act of 1988.

Dagnino, Renato (Ed.) Tecnología social. Ferramenta para construir outra sociedade (2º Edición, revisada y ampliada). Campinas, SP: Komedi. 2010.

Havlik, Jarmila. La tecnología y la discapacidad: Las tecnologías al servicio de las personas con discapacidad. (2000)

Indec. Censo Nacional de Población Hogares y Viviendas del año 2010.

López Cerezo J., Gómez González F. Apropiación social de la ciencia. Biblioteca Nueva–OEI. Madrid. 2008.

Toboso, M., Arnau, M. S. La discapacidad dentro del enfoque de capacidades y funcionamientos de Amartya Sen. Araucaria. Revista Iberoamericana de Filosofía, Política y Humanidades, Año 10, Nº 20, pp. 64-94, 2º semestre 2008.

Von Hippel, E. The sources of innovation, Oxford Univ. Press, New York, 1988.

Von Hippel, E. Democratizing Innovation, MIT Press, Cambridge, MA, 2005.

Winocur, R. Nuevas tecnologías y usuarios. La apropiación de las TIC en la vida cotidiana, Revista Telos, nº 73, octubre-diciembre 2007.

Zappalá, Daniel; Köppel, Andrea; Suchodolski, Miriam. Inclusión de tic en escuelas para alumnos con discapacidad intelectual. Octubre 2011.

Zappalá, Daniel; Köppel, Andrea; Suchodolski, Miriam. Inclusión de TIC en escuelas para alumnos con discapacidad motriz. Septiembre 2011.

Estrategias de apredizaje en procesos mediados por TIC: una experiencia con alumnos ingresantes

Tatiana Inés Gibelli

Universidad Nacional de Río Negro, Sede Atlántica
Viedma, Río Negro, Argentina
tgibelli@unrn.edu.ar

Resumen. Las tecnologías de la información y la comunicación (TIC) han ido produciendo grandes cambios en la sociedad, en particular, en el acceso al conocimiento y como consecuencia, en las formas de aprendizaje. Con el objetivo de indagar en el uso de estrategias de aprendizaje por parte de los alumno cuando el proceso es mediado por TIC se realizó una experiencia en un curso de matemática de primer año universitario. La misma consistió en un curso dictado en modalidad de aula extendida (blended learning) donde la enseñanza presencial se complementó con el uso de un entorno virtual implementado en plataforma Moodle. Esta experiencia incluyó actividades específicas para estimular el uso de estrategias y la autorregulación del aprendizaje. El análisis se centró en observar el impacto de la propuesta en el uso de estrategias. En este trabajo se presenta, en primer lugar, el marco teórico y la metodología de investigación propuesta, incluyendo las características de la experiencia llevada a cabo. Luego se exponen y analizan los principales resultados obtenidos como consecuencia de la implementación. Finalmente se proponen algunas conclusiones.

Palabras claves: Autorregulación, TIC, Estrategias, Aprendizaje, Matemática.

1 Introduction

En la sociedad actual, el acceso al conocimiento pasa, cada vez con mayor frecuencia, por las nuevas Tecnologías de la Información y la Comunicación (TIC), lo cual obliga a reconceptualizar los fines de la educación, y principalmente, la misma práctica docente. En nivel superior, las herramientas TIC permiten cambiar nuestras prácticas educativas, contribuyendo a la formación de los estudiantes universitarios, especialmente en la adquisición de las competencias necesarias para su futuro desempeño profesional. Una de las competencias a adquirir en el alumnado universitario, es la competencia digital, que implica aprender a gestionar la información que recibe así como el conocimiento que genera, es decir, aprender a buscar información, comunicarse, colaborar y participar [1]. Se trata de formar a los

alumnos para que sea capaces de entender los medios de comunicación actuales y saber utilizarlos [2]. Asimismo, la capacidad de autorregulación de los aprendizajes resulta esencial en cualquier tipo de estudios, especialmente, en nivel superior.

Teniendo en cuenta estas cuestiones, se propone una intervención para matemática universitaria de primer año, desarrollada en modalidad blended learning, complementando las clases presenciales con el uso de un entorno virtual implementado en plataforma Moodle. Se desarrolla una tarea de investigación en base a la implementación de dicha propuesta, cuyo principal objetivo es poder describir las características del aprendizaje de los alumnos en este entorno mediado por TIC. En este trabajo en particular se analizan el impacto de la propuesta sobre las estrategias de aprendizaje.

2 Marco teórico

El aprendizaje autorregulado es un tema de investigación relativamente reciente, con un abordaje cognitivo del aprendizaje, relacionándolo con formas de aprendizaje académico independientes y efectivas que implican metacognición, motivación intrínseca y acción estratégica [3]. Se define como *"un proceso activo en el cual los estudiantes establecen los objetivos que guían su aprendizaje intentando monitorizar, regular y controlar su cognición, motivación y comportamiento con la intención de alcanzarlos"* [4], y hace referencia a la capacidad del individuo de ajustar sus acciones y metas para conseguir los resultados deseados teniendo en cuenta los cambios en las condiciones ambientales [5]. Se concibe al estudiante como parte activa y fundamental del proceso de aprendizaje, centrada en la persona que aprende, y no solo en lo que aprende, sino y sobre todo en relación a cómo aprende [6].
Diversos autores han puesto el énfasis en analizar si es posible enseñar a autorregular el proceso de adquisición del conocimiento ([7], [8], entre otros). Varios de ellos concluyen que es necesario considerar el papel del adulto en el desarrollo de la autorregulación y particularmente, la estimulación para el desarrollo del aprendizaje autorregulado. Existen algunas investigaciones sobre intervenciones y modelos instruccionales diseñados con el objetivo de enseñar los procesos y las estrategias involucradas en el aprendizaje autorregulado [9]. Torrano y González-Torres plantean que los puntos en común en estas intervenciones son *"la enseñanza directa de estrategias, el modelado, la práctica guiada y autónoma de estrategias, la retroalimentación, la auto observación, el apoyo social y su retiro en el momento en que el estudiante ha alcanzado cierto grado de participación responsable y la autorreflexión"* [10].
Las TIC aplicadas a la educación muestran un gran potencial para el desarrollo de estrategias autorregulatorias del aprendizaje por parte de los estudiantes. En referencia a la enseñanza de la matemática específicamente, la integración de las TIC ofrece al estudiante la interacción y manipulación de contenidos y problemas matemáticos, permitiendo modificar condiciones, controlar variables y manipular fenómenos. Este hecho brinda al alumno, la capacidad de mejorar el pensamiento crítico y otras habilidades y procesos cognitivos superiores, motivando e involucrándolo en actividades de aprendizaje significativo.
Por otra parte, los ambientes de aprendizaje a distancia, cuyo soporte principal son las TIC, favorecen el seguimiento de metas personales, la libre navegación por los nodos

de información y resolución de diferentes situaciones problemáticas, de acuerdo con las diferencias individuales de los estudiantes [11]. Dichos entornos son una alternativa creativa a los soportes de aprendizaje más tradicionales para lograr la implicación de los procesos metacognitivos de los estudiantes en su aprendizaje [12], al mismo tiempo, que estimulan, mantienen y modelan ese proceso de autorregulación [13] que tan necesario se hace en el ámbito de la educación superior.
Respecto a las estrategias de aprendizaje, se adopta un punto de vista amplio integrando elementos afectivo-motivacionales de apoyo, metacognitivos y cognitivos, coincidiendo con Gargallo y colaboradores quienes las definen como *"el conjunto organizado, consciente e intencional de lo que hace el aprendiz para lograr con eficacia un objetivo de aprendizaje en un contexto social dado"* [14:p2]. Esta perspectiva es integradora y permite diseñar un mapa de estrategias poniendo énfasis en el uso de los diversos procedimientos y componentes que se movilizan para aprender, recogiendo elementos como: conciencia, intencionalidad, manejo de recursos diversos, autorregulación y vinculación al contexto.
En este marco se presenta una propuesta de intervención diseñada con el fin de promover la autorregulación del aprendizaje en estudiantes universitarios de matemática.

3 Metodología

3.1. Principios del diseño de la propuesta pedagógica

Algunas cuestiones que se consideraron como relevantes al momento del diseño de la propuesta pedagógica fueron:

- La necesidad de que las propuesta sea diseñada en contexto. Nuñez y colaboradores plantean que *"...la aplicabilidad real de las propuestas instruccionales realizadas no acaba de aportar los frutos deseados. Los avances de la investigación cognitiva, con frecuencia, no redundan en una mejora de la calidad del aprendizaje de los estudiantes, y no porque las prescripciones no sean epistemológicamente válidas sino porque tales propuestas han sido formuladas al margen del propio funcionamiento de los centros y todos los elementos que lo definen"* [15:p144].
- La enseñanza de la autorregulación junto con los contenidos disciplinares. Nuñez y colaboradores plantean la necesidad de considerar propuestas pedagógicas de inclusión transversal del trabajo de la autorregulación en él área de conocimiento específica [15].
- La instrucción en la autorregulación por andamiaje (scaffolding instruction). El suministro de apoyo social al alumno por parte de los profesores y de los compañeros a la vez que aprende las estrategias de autorregulación y la progresiva supresión del apoyo a medida que el estudiante sea más competente en su adquisición y desarrollo [10].

- Una propuesta de actividades que contemple la práctica autorreflexiva, brindando al alumno oportunidades para que autoobserve (self-monitoring) su aprendizaje. Este tipo de práctica se facilitará a través de la escritura de reflexiones personales sobre el propio aprendizaje a lo largo de todo el proceso.
- La inclusión de las TIC en dichas propuestas. La mayoría de las intervenciones corresponden a propuestas de tipo presencial, con tareas de papel y lápiz [16]. Sin embargo, las TIC se están convirtiendo en una de las variables críticas de los escenarios formativos.
- La evaluación del programas de intervención. En la revisión de distintos programas de intervención [16] plantean que tan sólo en algunos casos se evalúa la eficacia de las intervenciones y en muchos otros casos se proponen e implementan programas sin saber si realmente producen los resultados esperados.

3.2. Características de la propuesta implementada

La propuesta de intervención se orientó no sólo a que los estudiantes logren la comprensión de los contenidos específicos de la materia sino que, además, puedan mejorar sus conocimientos y habilidades en relación al uso de tecnologías y desarrollar la capacidad de autorregulación del aprendizaje. Dicha propuesta se abordó desde una perspectiva constructivista de orientación sociocultural de los procesos de enseñanza y aprendizaje [17].
La propuesta educativa fue diseñada para la materia Matemática I, correspondiente al primer año del plan de estudios de las carreras de Licenciatura en Administración Pública y Licenciatura en Gestión de Empresas Agropecuarias, que se dictan en el Centro Universitario Regional Zona Atlántica, de la Universidad Nacional del Comahue. El desarrollo de la materia se planificó en seis unidades de contenido denominadas unidades temáticas, que responden a núcleos conceptuales que forman parte del currículo de la asignatura. A su vez, las unidades temáticas se agrupan en tres bloques didácticos, en torno a los cuales se organiza el aprendizaje.
Cada bloque incluyó una secuencia de actividades de aprendizaje y de evaluación acordes a los objetivos de esta propuesta, partiendo de una gestión del proceso más guiada por parte del docente en el bloque inicial hasta permitir una mayor autonomía por parte del alumno en el bloque final. Se mencionan a continuación las instancias y tipo de trabajo propuestos en cada una:

1. Presentación de objetivos del bloque: el docente propone una guía del bloque y una agenda de trabajo recomendada. Este recurso que contribuye a una definición inicial compartida de los objetivos y las actividades del bloque que les permita orientar el aprendizaje y elaborar el plan de trabajo adecuado para llevarlo a cabo.
2. Desarrollo de unidades del bloque: se desarrollan en forma secuencial las dos unidades temáticas correspondientes al bloque. Cada unidad temática tiene una estructura estable con distintas líneas de trabajo que se realizan en forma paralela, mediante un conjunto recursos presenciales y virtuales, que se sintetizan a continuación:

Recursos		Descripción
Presenciales	Clases teóricas	Trabajo con cuestiones teóricas (conceptos, propiedades, modelos) de la asignatura.
	Clases prácticas	Realización de trabajos prácticos escritos sobre cuestiones prácticas (ejercitación, problemas).
	Clases de consultas	Espacios opcionales de consultas con los docentes sobre dudas acerca de los distintos temas.
Virtuales	Trabajos Grupales	Trabajo grupal que requiere del uso de recursos TIC y se envía a través del espacio virtual
	Diario de Aprendizaje	Actividades que permiten al alumno reflexionar sobre su proceso de aprendizaje.
	Cuestionarios de autoevaluación	Cuestionarios de corrección automática, que permiten al alumno auto-evaluar su comprensión de cada tema.

Tabla 1. Recursos con que se desarrolla cada unidad temática.

3. Cierre del bloque: los alumnos con la guía del docente elaboran una síntesis de los conocimientos adquiridos a lo largo del bloque. La finalidad es identificar los aspectos más relevantes y mostrar de forma explícita, las relaciones existentes entre ellos. El propósito es que los saberes puedan ser identificados por todos los implicados como el conocimiento construido y que se comparte; permitiendo además a los alumnos otra oportunidad para identificar y resolver dudas.
4. Evaluación de contenidos del bloque: se realizan actividades de evaluación para regular y valorar el aprendizaje alcanzado. Las mismas permiten tanto el aprendizaje de los conocimientos nucleares y relevantes del bloque como la recogida de información sobre el proceso de aprendizaje y de autorregulación de los alumnos.

3.3. Recolección y análisis de resultados

Para recoger datos sobre las estrategias de aprendizaje se ha aplicado el cuestionario CEVEAPEU (Cuestionario de Evaluación de las Estrategias de Aprendizaje de los Estudiantes Universitarios) desarrollado por Gargallo y colaboradores [14]. El instrumento fue validado con una muestra de estudiantes universitarios españoles, obteniéndo un coeficiente de fiabilidad (α de Cronbach) de 0.897. El cuestionario consta de 88 ítems en escala de tipo Likert con cinco opciones de respuesta: 1-Totalmente en desacuerdo, 2-En desacuerdo, 3-Ni de acuerdo ni en desacuerdo, 4-De acuerdo, 5-Totalmente de acuerdo. Los items se organizan en 25 estrategias agrupadas en dos escalas principales, una referida a estrategias afectivas, de apoyo y control y otra referida al procesamiento de la información. Para este trabajo las estrategias se agruparán en tres escalas: (i) estrategias motivacionales y afectivas, (ii)

estrategias metacognitivas, de apoyo y control del contexto y (iii) estrategias relacionadas con el manejo de la información.
El cuestionario CEVEAPEU fue aplicado en dos instancias: al inicio del curso (pre) donde se obtuvieron 82 respuestas y al final del curso (post) en que respondieron 27 alumnos. Con el objetivo de detectar si hay diferencia en la valoración de las estrategias en estas dos instancias, se consideraron las respuestas al cuestionario como muestras apareadas (es decir, respuestas del mismo alumno antes y después de la implementación). Bajo esta consideración, el número de casos disponibles para el análisis fue de 24 cuestionarios respondidos.
Se realizó una inferencia estadística para comparación de medias antes y después de la implementación, para indagar si las diferencias observadas eran estadísticamente significativas. Este análisis se realizó aplicando el test de hipótesis de comparación de medias para muestras apareadas utilizando la distribución t de Student. El análisis estadístico se realizó utilizando el software estadístico Infostat1.
Para indicar los cambios en la valoración de cada estrategia luego de la implementación, se utiliza la simbología: (↓) para indicar que disminuye, (↑) para indicar que aumenta y (=) para indicar que no hay cambios notables. Para la interpretación se considera que la diferencia de media es estadísticamente significativa de acuerdo al valor de probabilidad obtenido con la prueba bilateral (p-valor). En este caso vamos a utilizar la siguiente interpretación:

p-valor	Interpretación de diferencia
$p \leq 0{,}01$	Muy significativa
$0{,}01 < p \leq 0{,}05$	Significativa
$0{,}05 < p \leq 0{,}1$	Algo significativa
$p > 0{,}1$	No significativa

Tabla 2: Interpretación del valor de probabilidad en test de comparación de medias

4 Resultados

4.1. Estrategias motivacionales y afectivas

En la siguiente tabla se muestran los resultados del análisis estadístico correspondientes a las subescalas y estrategias de este grupo, considerando además los items que conforman cada estrategia:

Subescala de estrategias	Estrategia	Media post	Media pre	Media dif. (post-pre)	p-valor
Componentes	Motivación intrínseca (↑)	4,51	4,38	0,14	0,15

1 InfoStat es un software estadístico desarrollado por un equipo de trabajo conformado por docentes-investigadores de la Universidad Nacional de Córdoba, Argentina. (http://www.infostat.com.ar/)

internos	Atribuciones internas (=)	4,15	4,08	0,07	0,59
	Autoeficacia y expectativas (↓)	3,69	3,82	-0,14	0,26
	Inteligencia como modificable (↑)	4,19	4,04	0,15	0,35
	Valor de la tarea (=)	4,28	4,24	0,04	0,65
Componentes externos	Motivación extrínseca (=)	2,92	2,88	0,04	0,84
	Atribuciones externas (↓)	1,96	2,33	**-0,38**	**0,01**
Componentes afectivos	Estado físico y Anímico (=)	3,73	3,72	0,01	0,94
	Ansiedad (↓)	3,32	3,66	**-0,33**	**0,02**

Tabla 3: Comparación de medias en estrategias motivacionales y afectivas

En las estrategias motivacionales y afectivas se observa una disparidad en la valoración: en estrategias vinculadas a la motivación interna y auto-percepción los alumnos manifiestan un valoración positiva alta, mientras que las vinculadas a componentes internas presenta una valoración media-baja. En componentes afectivas, se manifiesta un buen estado físico-anímico, pero con un nivel alto de ansiedad.

Respecto a los cambio observados en este grupo de estrategias se pueden mencionar una disminución muy significativa (0,38 pts) de las atribuciones externas y disminución significativa de la ansiedad (de 0,33 pts). Se observan además modificaciones en valoración que no son significativas: aumento de la motivación intrínseca (de 0,14 pts) y de la concepción de la inteligencia como modificable (de 0,15) y disminución de la valoración de la autoeficacia y expectativas (de 0,14 pts).

4.2. Estrategias metacognitivas y de control

En la siguiente tabla se muestran los resultados del test de hipótesis para diferencia de medias en la ponderación de estrategias metacognitivas y de control:

Subescala de estrategias	Estrategia	Media post	Media pre	Media dif. (post-pre)	p-valor
Estrategias de organización	Conocimiento de objetivos y criterios de evaluación (=)	3,46	3,42	0,04	0,77
	Planificación (↑)	3,41	3,16	**0,25**	0,09
Estrategias de auto-control	Autoevaluación (=)	3,65	3,61	0,04	0,56
	Control y Autorregulación (=)	3,91	3,91	0	1
Control del contexto e interacción social	Control del contexto (=)	4	3,91	0,09	0,5
	Habilidades de interacción social y aprendizaje con compañeros (↑)	4,01	3,91	0,1	0,52

Tabla 4: Comparación de medias en estrategias metacognitivas

Se puede observar una valoración medio-alta de todas las estrategias de esta escala. Tienen menor ponderación las estrategias de organización (conocimiento de objetivos y planificación) seguidas por las estrategias de auto-control (autoevaluación y control-

autorregulación), mientras que las estrategias vinculadas a control de contexto e interacción social presentan los mayores puntajes.

Como principal cambio en este grupo de estrategias metacognitivas y de control se puede mencionar el incremento en la estrategias de planificación, que resulta algo significativo. Otra de las estrategias de este grupo que presenta un leve incremento, que no lleva ser significativo son las habilidades de interacción social y aprendizaje con compañeros. El resto de las estrategias de este grupo no presentan cambios notables.

4.3. Estrategias de manejo de la información

En la tabla siguiente se muestran los resultados del análisis estadístico realizado para este grupo de estrategias:

Subescala de estrategias	Estrategia	Media post	Media pre	Media dif. (post-pre)	p-valor
Búsqueda y selección de la información	Conocimiento de fuentes y búsqueda de información (=)	3,42	3,38	0,04	0,79
	Selección de información (↑)	3,52	3,4	0,13	0,32
Incorporación de la información	Adquisición de información (=)	3,63	3,56	0,07	0,65
	Organización de información (=)	3,8	3,73	0,07	0,57
Procesamiento de la información	Elaboración de información (=)	4,15	4,19	-0,04	0,65
	Personalización y creatividad (↑)	3,51	3,31	0,2	0,23
Almacenamiento de la información	Simple repetición (↑)	2,81	2,58	0,23	0,18
	Memorización. Uso de recurso mnemotécnicos (↑)	3,08	2,87	0,21	0,39
Uso de la información	Manejo de recursos para usar la información (↑)	3,96	3,79	0,17	0,25
	Transferencia de la información (↑)	3,9	3,67	0,24	0,11

Tabla 5: Comparación de medias en estrategias de manejo de la información

Todas las estrategias consideradas en este grupo tienen una valoración medio-alta por parte de los alumnos. Las estrategias de este grupo que tienen mayor valoración son las de uso de la información, mientras que las estrategias de almacenamiento de la información, son las de menor valoración.

Respecto a los cambios, en general, puede decirse que la mayor parte de las estrategias vinculadas al manejo de la información presentan un leve aumento (de 0,2 pts aprox.) que no resulta significativo, mientras que algunas se mantienen sin cambios notables. Más específicamente, las estrategias correspondientes a subescalas de almacenamiento y uso de la información presentan un incremento, así como también se obervan leves incrementos en estrategia de personalización y creatividad (0,2 pts) y en estrategia de selección de la información (0,13 pts).

5 Conclusiones

El análisis realizado permite hacer una descripción del uso que manifiestan los estudiantes respecto a las distintas estrategias. Los alumnos presentan un uso adecuado (valoración medio-alta) de las estrategias metacognitivas, de apoyo social y control del contexto, así como de las distintas estrategias vinculadas al manejo de la información. En cuanto a motivación, se observa una gran influencia de los componentes internos y de auto-percepción (valoración alta) mientras que los componentes externos tienen baja influencia, en especial la atribución de rendimiento a agentes externos. Por otro lado si bien expresan un estado físico anímico adecuado para el estudio se observan altos niveles de ansiedad.

En cuanto al impacto de la propuesta, teniendo en cuenta los resultados mencionados anteriormente, se puede concluir que los mayores cambios se observan en las estrategias motivacionales y afectivas: disminución muy significativa de las atribuciones externas y disminución significativa de la ansiedad. En estrategias metacognitivas y de control se destacan un incremento en la estrategias de planificación y en habilidades de interacción social y aprendizaje con compañeros. Respecto a las estrategias vinculadas al manejo de la información se mantienen sin cambios notables, observándose una leve tendencia de aumento. Cabe mencionar que los cambios observados son pequeños ya que en ningún caso la variación promedio ha superado los 0,4 pts en un rango de 1 a 4. Lo que se puede observar es una tendencia de modificación. No debe olvidarse que se trata de un lapso muy breve (un cuatrimestre) el que se considera. Sería de esperar que un trabajo a largo plazo en el mismo sentido permitiese el logro de cambios más notorios.

Respecto a los recursos utilizados en la propuesta, Diario de Aprendizaje es el que parece haber tenido el mayor impacto pues los cambios como disminución de las atribuciones externas y de la ansiedad y el incremento en la estrategias de planificación podrían atribuirse al tipo de actividades propuestas en utilizando este recurso. El aumento en habilidades de interacción social y aprendizaje con compañeros podría deberse a los Trabajo Grupales propuestos. Es de destacar que estos dos recursos mencionados son de tipo virtual, o sea están mediados por TIC.

Podría afirmarse entonces que la propuesta diseñada, a pesar de lo acotada en el tiempo, muestra un impacto positivo en el uso de estrategias y la autorregulación del aprendizaje en los alumnos: Asimismo, los recursos TIC parecen tener un rol importante en este proceso. Sin embargo, se plantea la necesidad de que este tipo de

trabajo tenga una continuidad en el tiempo, quizá involucrando al alumno en actividades de este tipo en otras materias de la carrera.

References

1. Monereo, C. Pisa como excusa. Repensar la evaluación para cambiar la enseñanza (coord.) Barcelona. Graó. (2009)
2. Ortega Carrillo, J.A. Los medios didácticos t su tenología. En "Didáctica General. La práctica de la enseñanza en la Educación Infantil, Primaria y Secundaria" de Agustín de la Herrán Gascón y Paredes Labra, J. (coord). Madrid: McGraw-Hill. (2008)
3. Perry, N.E. Introduction: Using qualitative methods to enrich understandings of self-regulated learning. Educational Psychologist, 37(1), pp. 1-3. (2002)
4. Rosário, P. Estudar o Estudar: As (Des)venturas do Testas. Porto: Porto Editora. (2004)
5. Zeidner, M., Boekaerts, M. y Pintrich, P. Self-regulation: Directions for future research. In M. Boekaerts, P. Pintrich & M. Zeidner (Eds.). Handbook of self-regulation (pp. 749-768). San Diego: Academic Press. (2000)
6. Cochram-Smith, M. Teaching quality matters. Journal of Teacher Education, 54 (2), pp. 95-98. (2003)
7. Pozo, J.I y Monereo, C. El aprendizaje estratégico. Madrid: Santillana. (2002)
8. Simón, M., Márquez, C. y Sanmartí, N. La evaluación como proceso de autorregulación: diez años después. Alambique, 48, pp 32-41. (2006)
9. Schunk, D.H., y Zimmerman, B.J. (Eds.) Self-regulated learning: From teaching to self-reflective practice. New York: Guilford Press. (1998)
10. Torrano, F. y González-Torres, M.C. El aprendizaje autorregulado: Presente y futuro de la investigación. Revista Electrónica de Investigación Psicoeducativa, 2 (1), 1-34. (2004)
11. Jacobson, M., y Archodidou, A. The design of hypermedia tools for learning: Fostering conceptual change and transfer of complex scientific knowledge. Journal of the Learning Sciences, 9(2), pp. 145- 199. (2000)
12. Zimmerman, B. J. y Tsikalas, K. E. Can Computer-Based Learning Environments (CBLEs) Be Used as Self-Regulatory Tools to Enhance Learning?. Educational Psychologist, 40(4), pp. 267–271. (2005)
13. Coll, C. Psicología de la educación y prácticas educativas mediadas por las tecnologías de la información y de la comunicación: una mirada constructivista. Sinéctica, 25, 1-24. (2004)
14. Gargallo, B., Suárez-Rodriguez, J.M. y Pérez-Perez,C. El cuestionario CEVEAPEU para la evaluación de las estrategias de aprendizaje de los estudiantes universitarios. RELIEVE, 15(2), 1-31. (2009)
15. Núñez, J.C., Solano, P., González-Pienda, J.A. y Rosário, P. . El aprendizaje autorregulado como medio y meta de la educación. Infocop, 3 (21). (2006)
16. Cerezo, R., Núñez, J.C., Fernández, E., Suárez-Fernández, N. y Tuero E. Programas de intervención para la mejora de las competencias de aprendizaje autorregulado en educación superior. Revista Perspectiva Educacional, Vol 50, N° 1, pp. 1-30. (2011).
17. Azevedo, R. Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. Educational Psychologist, 40(4), pp 199-209. (2005)

Epistemological obstacles in the learning process of Numeral Systems

Marcia Mac Gaul[1], María Laura Massé Palermo[1] and Paola del Olmo[1]

[1]Research Council of the National University of Salta, Av. Bolivia 5150, 4400 Salta, Argentina
mmacgaul@cidia.unsa.edu.ar, mlmassep@cidia.unsa.edu.ar, pdelolmo@unsa.edu.ar

Abstract. This study has two main research objectives. Its main one is to analyze the origin of a systematic error, made by first year students of the Computer Sciences Degree at the National University of Salta. The systematic error arises when the students are dealing with Numeral Systems. From there, a second objective is to develop teaching procedures leading to recreate the error in a conceptual reconstruction process, within a playful environment. The theoretical framework used is the Brousseaunian concept of epistemological obstacles in the learning process. We use a sample of 152 students. In this article, we propose a strategy that allows real processes of error reconstruction and construction, in order to enable real epistemological progress. Resources within a virtual environment are presented as play and interactive alternatives with the aim of reinforcing the concepts studied in the classroom: videos, educational software and a computer game specifically developed for Numeral Systems.

Key words: Epistemological obstacles, Numeral Systems, Error Construction and Reconstruction, Teaching procedures, Liaison.

1 Introduction

This study has two main research objectives. The first one is focused on the origin of a systematic error made by first year students of the Computer Sciences Degree at the National University of Salta, when dealing with mathematical concepts that have been taught in different theoretical and conceptual domains. The second one is to develop an innovative teaching procedure, seeking the reconstructing of the systematic error through a conceptual revision process, within a playful environment.

Activities were carried out within the framework of the CIUNSa Research Project N° 1865/3 called *Virtual Environments for the between Secondary School and Exact Sciences University Degrees* which has been accredited by the Research Council of the National University of Salta. The timeframe for this project is from 2010 to 2013. Students in their last year of Secondary School, who are interested in studying either an university degree in Math, Chemistry or Computer Sciences, take part in this project. In a second stage, a follow-up system is used for students who actually go on to University. It is therefore of interest to detect conceptual problematic foci on each discipline within the University environment, in order to guide the Liaison activities based on these results.

The innovative teaching procedure is applied then to first year students at the University and to Secondary School students taking part in the Liaison project.

The inter-disciplinary team involved in the Project upholds that University teaching and research are complementary and mutually reinforcing activities [1]. Some authors of this study, in their role as teachers, use the teaching procedures developed from their research, in their classrooms. This kind of research improves the teaching process, involving the teachers in a double role, as teachers and researchers, who simultaneously perform at three different levels, Empiric, Theoretical and Meta-theoretical.

2 Application context

The IT subject courses in the first year of the Degree in Systems Analysis (LAS) and University Technical Degree in Programming (TUP), given by the Faculty of Exact Sciences of the National University of Salta, have really high matriculation numbers. In general, the students have recently finished Secondary School. It is not always the case that the students have the required cognitive ability to carry out a certain level of abstraction that enables them to keep the pace in the first year. Neither they have developed study habits or strategies that allow them to settle themselves in a Higher Education System. The subject course called Elements of Programming is part of both the LAS and TUP's syllabi. It is taught in the first semester of the first year, and therefore it is the first IT subject course the students take. The subject's content can be divided in three:

- Initial Programming Concepts, with a focus in algorithms design.
- Basic computing elements that have base on Applied Mathematics: Numeral Systems and Boolean Algebra.
- Complementary and introductory content leading to computing literacy.

3 Objectives

The main aim is to analyze the origin of a systematic error made by the LAS and TUP first year students, when dealing with Numeral Systems. From there we aim to design an innovative teaching procedure, leading to reconstruct the aforementioned error.

The theoretical framework answers to the Brousseaunian concept of epistemological obstacles in the learning process. Due to the nature of the subject contents, which is usually known by the students in other contextual domains -such as counting and solving basic operations within the Decimal System-, we resort to a playful environment for the conceptualization and practice on this and other numeral systems.

3.1 Obstacles of epistemological and didactic origin

Mathematics provides several examples of obstacles that result in different kind of errors. The knowledge of the Decimal System as an obstacle to the learning of other Numeral Systems, which are especially useful for Computer Sciences, such as Binary, Octal and Hexadecimal Systems.

In the Decimal System, the numbers have a specific name that identifies them. Let say, for example, thirteen, which in mathematical notation is $(13)_{10}$ or just 13, when the sub-index that indicates the system's base is omitted. In a different numeral system, such as the Octal System, the number 13 is represented by $(15)_8$. The correct denomination of this octal number is one-five, reading the number's consecutive characters, and not fifteen, as a Decimal System reading would mistakenly suggest.

It can be noticed the need for abstraction that is required to keep stable, not only the quantity, but also the oral and written representation of numbers when they are expressed in different numeral systems. The abstraction reference is essential for this analysis. Guillermo Simari in [2], says:

"Our graduates must have the ability to think in different levels of abstraction and this ability is difficult to acquire, it requires time and practice." And adds: *"The depth and complexity in the change of perspective make it necessary to approach the teaching of these abilities in an early stage, in order to allow enough time for the indispensable comprehension and cognitive maturity process."*

In addition to representational difficulties, there are others whose origin lay in real conflicts that originate in previous knowledge. According to [3]:

> *"... to describe knowledge, understand its usage; explain which advantages it provides compared to previous uses, to which social practices is related, with which techniques and, if it is possible, with which mathematical conditions; to indicate these conceptions in relation to other possible conceptions, especially those that succeeded it, in order to understand its limitations, its difficulties and finally the causes for this conceptions' failure, but at the same time the reasons for a balance that seems to have lasted a long enough time; to search possible reappearances, unexpected environments, if not under its initial appearance, at least under similar forms, and the reasons for that."*

Regarding the balance that lasts over the time, few pieces of knowledge are more general, with such a practical use, and so socially related, as that which states that "any number ending in an even digit, is an even number itself". When venturing into new numeral systems, the validity of this statement is challenged. For example, we can see this in a numeral system whose set of characters has an odd number of elements. The set of characters is $C= \{0, 1, 2, 3, 4\}$. The base is $\beta = (10)_5 \equiv (5)_{10}$.

The following chart shows the succession of the first decimal system's positive integers, the numeral system in base 5 and a third one equivalent to the latter, and whose characters are abstract. It's not difficult to notice that the decimal system's even numbers have their corresponding even equivalents in the numeral system in base 5. Some of them, such as 11 (YY) and 13 (YW) do not follow the statement above, despite them definitely being even numbers.

Chart 1. Decimal counting and systems with an odd base.

Base	Succession of positive integers													
In base 10:	0	1	2	3	4	5	6	7	8	9	10	11	12	…
In base 5:	0	1	2	3	4	10	11	12	13	14	20	21	22	…
In base YZ: C = {Z,Y,X,W,V}	Z	Y	X	W	V	YZ	YY	YX	YW	YV	XZ	XY	XW	…

4 Method

The error which is being under study pertains to the subject Numeral Systems, given in Unit No. 6 of the course called Elements of Programming.

It is worth mentioning that in the first part of the course, basic programming concepts are taught. It comprises the Computing problems resolution, Basic algorithms and one-dimensional and bi-dimensional indexed variables. This is important because the algorithmic approach is cross-linked to the rest of the subjects that are taught, trying to encourage abstraction abilities in different problem situations, which are typical of Computer Sciences. Therefore, before studying the Numeral Systems, the student learns to design algorithms in which they frequently uses different algorithms such as the separation of a number's digits, composition of a number from the weighted sum of its digits, parity recognition using the MOD function, among others.

After the units 6 and 7 are taught, students are evaluated through a written test on Numeral Systems and Boolean Algebra. It comprises multiple-choice exercises: 5 for the former subject and 10 for the latter. Each exercise has four answers to choose from. And only one is correct. The exercise on which this study is focused is:

> *Let C = {Z, Y, X, W, V} be a set of characters of a numeral system in base β = YZ. Please indicate the smallest even natural number that has two significant digits.*

The four choices given are: a) YZ, b) ZY, c) YY (correct answer) and d) None of the above. The two first wrong choices answer to the following criteria:

A) This is a choice that fits the expected error. If we translate the presented Numeral System to a system in base 5, the number YZ is equivalent to $(10)_5$. Indeed, if we only analyze the written representation $(10)_5$ or its oral representation *one-zero*, we can conclude that it is an even number because it ends in an even figure. It is clear that $(10)_5$ is an odd number due to the following: a) in the number line, the number $(10)_5$ is between two even numbers, b) the decimal number equivalent to $(10)_5$ is 5, which is an odd number, c) $(10)_5$ cannot be written as *2*n*.

B) This is an incorrect answer that does not follow any of the conditions asked in the exercise. The number ZY is equivalent to $(01)_5$, or $(1)_5$ because zero is not a significant digit in that position, and therefore, it is not a number with two digits. Additionally, 1 is an odd number.

5 Materials and Methods used for the Teaching Procedure

For a few years now, the department has been working with a *B-Learning* method, using a virtual course called *Elements of Programming*, designed on a Moodle platform. This course's goal is to open an alternative meeting space through which the students could establish other communication channels with their peers and teachers, as well as find different activities and material that may help to reinforce their learning process.

The virtual classroom's subjects are organized in blocks of topics, one for each unit of the syllabus. The blocks corresponding to the different units of the syllabus have the theoretical and practical material required to attend classes on campus, a virtual self-evaluation questionnaire and additional material related to the unit's specific subjects. In the block corresponding to Numeral Systems, as seen in Figure No 1, the students can access a series of videos, games and software that provide alternative play and interactive activities to strengthen the knowledge acquired in the on-campus classroom. The students are given the possibility of designing and implementing in a collaborative way the algorithms for conversions between numeral systems, by means of a wiki, and using a pseudocode.

Although all the resources put forward help to prop up the Numeral Systems' concepts, it is of special interest for this study to focus on those resources aimed at re-conceptualizing the error being studied. Most of the resources that are made available have an interactive and play-orientated approach. The use of this kind of resources aims at making available to the students alternatives that are different, motivating and entertaining, whilst educational.

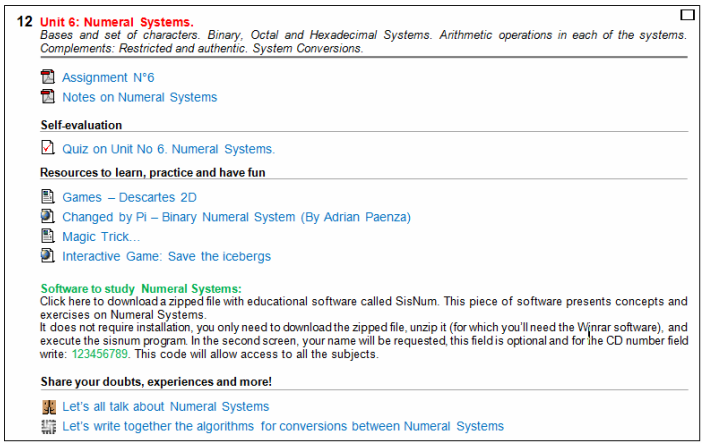


Fig. 1. Numeral Systems Block of the "Elements of Programming" virtual course.

One of these resources is the game "Save the icebergs". This game was developed by the department using Alice 3.1. [http://www.alice.org/]. The objective of this game is that through the solving of riddles presented to the player, he/she can help to save the icebergs from total destruction. The game takes place in a planet where its

inhabitants have three fingers and therefore it is natural to count in a numeral system in base 3.

Some of the questions are:

- If the first even number is 2, which is the following one? Answer: 11. Once that it is clear that the following even number is 11, the following questions are asked:
- Given 11+2, is the result an even number? Press 1 for YES or 0 for NO. Answer: YES
- If $11)_3$ equals $4)_{10}$ then, does every even number in this system end with a 1? Press 1 for YES or 0 for NO. Answer: NO

The game is structured in two steps for each riddle, first a challenge is presented to the player, if he/she can solve it, then he/she can move forward towards the solving of other riddles. The second step is when a player fails to answer correctly and, as a consequence of his/her error, an iceberg disappears. It is then when the main character of the game presents to the player the possibility to recover the lost iceberg, if together they try to find the right answer. To this end, the character uses different resources, such as his fingers and some ice rocks, to count. The possibility of reasoning "together", the player and the game's character, aims at showing the correct logical reasoning to achieve the right answer. Finally, the player has a new opportunity to answer the riddle.

In order to show an example, the actions carried out in the game as a result of a "wrong" answer in the first question (it has inverted commas because the student believes there is no error) are described. In this case, the character counts with his fingers, starting with 1 and up to 11 (one-one). Simultaneously, below the character, the ice stones are grouped together until there are 4 in total. At this moment, the player is asked again. If the players gives the correct answer, the lost iceberg is recovered and the games continues, otherwise the player loses the game. Figure 2 show a screenshot of the game.

Fig. 2. Screenshot 1 - "Save the icebergs" game.

6 Results

The test was carried out on a sample of 152 students who took the course Elements of Programming in the first semester of 2013. For the purposes of this study, only the results for the block corresponding to Numeral Systems are considered. Table 2 shows the distribution of answers given by students, divided between those who passed the block and those who failed. In addition, it distinguishes –out of the wrong

alternatives– which are those selected by the students.

Table 2. Distribution of answers to the exercise set out.

Answers	Passed			Failed			Total/Percentage
Correct (Alternative C)	40			4			44
Percentage	36%			10%			29%
Incorrect							
Alternative	A	B	D	A	B	D	
	49	4	17	22	6	10	
Total Incorrect	70			38			108
Percentage	70%	6%	24%	58%	16%	26%	71%
Total Answers	110			42			152
Percentage	72%			28%			100%

From the results we can infer a first approximation to the identification of the error as an epistemological obstacle.

- 72% of the students passed the Numeral Systems block, with a minimum of 3 out of 5 correct exercises. The four exercises pose operations or conversions in numeral systems, whose solutions are closely linked to the application of algorithms of arithmetical calculus, which demand an abstraction level that is lower than the problem under study.
- Of the 44 students who passed, it is not surprising that 36% responded correctly to the problem, which is much higher than the 10% that the answer correct in spite of not having passed the block.
- The highest percentage observed when answering incorrectly, corresponds in both cases to the alternative A. 70% among those who passed and 58% among those who failed . This result is coherent with the conjecture of the error that applies the parity rule within the Decimal System's domain.
- The less frequent error belongs to alternative B, which is a distracting alternative. It is natural that there is a lower percentage of choice, 6%, from students that passed, compared to the 16% obtained from students that failed.
- The alternative D has a medium level of frequency. This fact can be explained on the basis that the student who is in doubt tends to choose the answer that denies the other three alternatives.

The results, as previously mentioned, allow for a first conjecture. In the following, we analyzed if it is an epistemological obstacle that generates this error. Brousseau defines it as that which corresponds to a body of knowledge that in a previous time had a progressive direction and allowed access to a certain level of knowledge, but on attempting to widen the domain of such knowledge, difficulties arise. Therefore, the obstacle appears as a hindrance for a deeper and broader comprehension.

In order to specify in detail the idea of obstacle, the conditions that must fulfilled, according to Susana Quaranta and Maria Emilia Wolman, 1995 [4], are stated and applied to this study in table 3, below:

Table 3. Conditions to characterize an epistemological obstacle.

Condition	Characterization of the case under study
The errors correspond to conceptions that support them, which maybe either general or particular. The errors are supported by implicit theories that are coherent and consistent for the student.	There is a double difficulty: on the one hand, the error originates from a more specific conception (the specific fact of having always worked with a Decimal System), and at the same time there is an awareness that this knowledge is of a general nature. In any case, it is clear that the problem lies in the existence of "incomplete" or "relative" knowledge and not in the fact that there is a lack of knowledge, that is, always from the teacher's point of view.
An obstacle has a domain of validity and effectiveness. It exists in an area where that knowledge is correct. Said knowledge is effective within certain boundaries.	It is clear that the boundaries of the domain of knowledge are the problem to be looked into, when solving situations outside that domain. The arising question is the following: to what extent could we talk of a correct piece of knowledge if the same does not behave as such outside the domain, which is a part of a larger whole? In other words, assuming that the Decimal system is adopted, but that it coexists with as many as we would like to define, which is to say that there is an awareness of operating within a part of a whole, wouldn't it be reasonable to adjust to rules whose validity can be guaranteed for the whole and not limited to a part of the domain, which, inevitably, will become wrong for domains that are close, but different? Naturally, these questions are aimed to the teachers.
Brousseau states, "if out of that context, it generates false answers". There are problems where the piece of knowledge held by the person appears to be relevant, but turns out to be false, ineffective and a source of errors. The student is not aware that his/her knowledge is of a local nature, does not know the boundaries beyond which it loses its validity, coherence and validity.	The student is not aware of the boundaries of the domain of his/her knowledge. This is the case, because said knowledge was presented with a kind of universal validity. In this case, the statement "any number ending in an even digit, is an even number itself", gives the word any a strong semantic meaning, which is very difficult to change when assessing the validity of a proposition.
"… this piece of knowledge resists the contradictions it faces and the rooting of a better piece of knowledge. It is not enough to have a better piece of knowledge for the previous one to disappear, this is what differentiates the overcoming of obstacles from Piagetian	The results obtained are an eloquent proof of the resistance exerted by the previous piece of knowledge, when faced with a jeopardizing situation presented by the new piece of knowledge. 71% of the students that had to reconsider their previous concepts, not so much as incorrect but as unfinished before the new piece of knowledge, when tested on their reflections about the behavior of rules formerly

accommodation. It is essential to identify it and to contrast it against the new piece of knowledge. In order to establish its obstacle nature, it is essential to prove and explain the resistance to reject one piece of knowledge and to the rooting of a more accurate one.	known, gave way to the local knowledge mentioned above.
Obstacles do not disappear suddenly. They continue to appear, and keep on arising long after having been conscientiously rejected.	There are no data on the persistence, or not, of the error under this study.

Resuming the subject of obstacles that originate at school, in a conference in the UQAM, Canada, on January 21, 1988, Brousseau stated the following on the link between obstacles and the teaching-learning contract [5]:

"The student's place in the teaching-learning has been reclaimed by different disciplines -psychoanalysis, psychology, pedagogy, etc.- as the place of "reality"". Genetic epistemology has offered the most serious arguments and those closer to knowledge, but other studies are necessary to use its contributions. It is frequent that the student's mistakes are read by the teacher as an inability to reason in general or, at least as a logical error: within a broader teaching-learning contract, the teacher takes charge of the representations, of the direction of knowledge. But, within narrower conditions, he/she is only led to point where the student's answer contradicts previous knowledge, carefully avoiding any diagnosis on the cause of the error. .... This teaching-learning contract provided the teacher the safest defense: He/she only takes responsibility of the knowledge already known within his/her own domain. It is enough for the teacher to set an axiomatic order and then demand axioms as answers."

If the teacher does not enquire about the student's representations, because of his/her own disciplinary and/or pedagogical shortages, there is a risk of upholding teaching procedures that are not the correct ones to teach Numeral Systems. According to Delia Lerner and Patricia Sadovsky [6]:

"The hypothesis according to which the numeral writing arises from its correlation to oral numeracy, leads children to produce non-conventional notations. Why? Because, unlike written numeracy, oral numeracy is not positional".

An example of the statement above, is when the child writes four thousand seven hundred and five, just as he says it: 4 1000 700 5.

In the problem treated in this study, the number one-one, which is the right answer, is presented in a type of representation in which it is difficult to identify an even number. Only through the counting of integers and the acknowledging of the alternation between even and odd numbers, on the number line, or through the conversion of $(11)_5$ to the Decimal System, by applying a weighted sum of powers of the base, or in other words, by resorting to the positional nature of the numeral systems, it is possible to recognize one-one as the equivalent to 6 in the Decimal System, which is an even number in this system and therefore even number in any

other system. There is research suggesting that the child apprehends the numeral system, not from the intrinsic characteristics of the system, that is, from its positional condition, but rather from conceptualizations that he/she has on quantities, knowledge that comes from the oral world, before the written representation.

7 Some conclusions

The conclusion is therefore, that when revisiting this body of knowledge in an introductory university course, or within a Liaison experience with future University students, special attention must be put on written representations and representational discourse of numbers. Both types of representations coexist in the numeral system in base 10, which the student uses all throughout school. Incorporating the Binary, Octal, Hexadecimal systems or any other, including systems such as the one presented here with generic characters, requires a higher level of abstraction in order to maintain stable those previous pieces of knowledge and revisiting the validity of the rules in this new domain.

The present landscape, in which scientific-technological University Degrees show a strong negative tendency with regards of student retention and persistence rates, challenges the teacher to revisit his/her practices, searching for teaching procedures that allow real processes for construction and deconstruction of errors, as a real source of epistemological progress. Likewise, the University, continuing with the educational policy of universal admission, should review its support and guidance strategies available for students, who experience complex and heterogeneous learning paths. To assume the challenge of rising numbers of admitted students means to assume the problem of the diversity and inequality of previous knowledge with which they come with when entering University, which is the result of a fragmented education system. We are convinced that Liaison activities can be a very useful tool. Additionally, the University should promote learning environments in which that previous body of knowledge is challenged, directed towards a process of conceptual change. Another fundamental aspects to have in mind are the students' interests, who are just starting to solidify their vocation towards Computer Sciences. In both cases it is essential to approach the syllabus' contents by means of activities that are enjoyable, creative and that require team work.

References

1. Mac Gaul, M., López, M. F.: Sistemas de Numeración: una metodología de enseñanza basada en el enfoque algorítmico. En VI Congreso de Tecnología en Educación y Educación en Tecnología – TE&ET. ISBN 978-987-633-072-5.2011.
2. Simari, G.: Los fundamentos computacionales como parte de las ciencias básicas en las terminales de la disciplina Informática. En VIII Congreso de Tecnología en Educación y Educación en Tecnología – TE&ET. 2013.

3. Brousseau G. : Obstacles épistémologiques, conflicts socio-cognitif set ingénierie didactique. En: Bodnarz N., Garnier C. (editores). Les obstacles épistémologiques et le conflit socio-cognitif. Constructiondessavoirs (obstacles et conflits)". 1989.
4. Wolman, S., Quaranta, M. E.: Tras las huellas del "h"error. Piaget y Brousseau focalizando los errores en los procesos cognitivos y didácticos. Documentos de trabajo 13. Buenos Aires, Instituto de Investigaciones en Ciencias de la Educación, Facultad de Filosofía y Letras, Universidad de Buenos Aires. 1995.
5. Brousseau G.: Los diferentes roles del maestro. En Didáctica de matemáticas. Aportes y reflexiones. Parra, C. y Saiz, I. (comp.). Ed. Paidós. (pp 65 a 94). 1994.
6. Lerner, D., Sadovsky, P.: El Sistema de Numeración: un problema didáctico. En Didáctica de matemáticas. Aportes y reflexiones. Parra, C. y Saiz, I. (comp.). Ed. Paidós. (pp 95 a 182). 1994.

Avances en el diseño de una herramienta de autor para la creación de actividades educativas basadas en realidad aumentada.

Lucrecia Moralejo[1, 2] , Cecilia Sanz[2], Patricia Pesado[2,3], Sandra Baldassarri[4].

[1]Becaria UNLP (Universidad Nacional de La Plata), Buenos Aires, Argentina.
[2] III-LIDI, Facultad de Informática, UNLP, La Plata, Buenos Aires, Argentina.
[3]CIC (Comisión de Investigaciones Científicas de la Pcia. de Buenos Aires), Argentina.
[4]Grupo de Informática Gráfica Avanzada (GIGA), Universidad de Zaragoza, Zaragoza, España.
{lmoralejo, csanz, ppesado}@lidi.info.unlp.edu.ar, sandra@unizar.es

Abstract. En este artículo se presentan los avances logrados en el diseño de una herramienta de autor, llamada AuthorAR, orientada a la creación de actividades educativas basadas en realidad aumentada (RA). AuthorAR permite generar actividades de exploración y de estructuración de frases, que pueden favorecer procesos de adquisición del lenguaje y de entrenamiento de la comunicación, por lo que se hará referencia a las posibilidades que ofrece en este sentido. Se presentan aquí: una descripción de esta herramienta de autor, una revisión de antecedentes en la temática y la propuesta de evolución de este proyecto, con los primeros resultados obtenidos y las conclusiones arribadas.

Keywords: Herramientas de autor; Realidad Aumentada, Tecnología Educativa.

1 Introducción

La realidad aumentada (RA) es una tecnología que está creciendo rápidamente, y ha empezado a integrarse a algunas de nuestras actividades. Complementa la percepción e interacción con el mundo real y permite al usuario estar en un entorno real aumentado, con información adicional generada por el ordenador.

Varios son los autores que remarcan el beneficio de la RA en el ámbito de la educación, y en particular, en la educación especial. Entre ellos, Lin y Chao en [1], remarcan que la RA puede ser aplicada al aprendizaje asistido por computadora, que permite diseñar materiales educativos atractivos, y que al mismo tiempo, puedan ser utilizados en situaciones que resulten beneficiosas para estudiantes con necesidades especiales.

En la sección 2 de este artículo, se hace una breve revisión de los conceptos básicos de realidad aumentada, y se muestran algunos antecedentes de aplicación de RA en educación en general, y para personas con algún tipo de discapacidad, en particular. En la sección 3, se presenta la motivación para el desarrollo de una

herramienta de autor que permita la creación de actividades educativas basadas en RA, y en la sección 4, se describe AuthorAR y sus posibilidades para docentes y alumnos, incluyendo algunos ejemplos de actividades para los escenarios de la educación especial. Por último, se presentan algunas conclusiones y las líneas de trabajo futuras.

2 Antecedentes sobre Realidad Aumentada

La realidad aumentada (RA) agrega información sintética a un escenario real. Algunos la definen como un caso especial de realidad virtual (RV), otros como algo más general, y ven a RV como un caso especial de RA.

Paul Milgram y Fumio Kishino [2] definieron en 1994 el "Reality-Virtuality Continium" como un continuo que va desde el "entorno real" hasta el "entorno virtual". Al área comprendida entre los dos extremos, donde se combinan lo real y lo virtual, la denominaron "Realidad Mezclada" (Figura 1).

Fig. 1. Definición de realidad aumentada por Milgram y Kishino

De esta forma, estos autores distinguen entre una "Realidad Aumentada", en la que se incorporan elementos virtuales a un entorno real, y la "Virtualidad Aumentada", en la que se incorporan elementos reales a un entorno virtual.

En la definición del autor Ronald Azuma, un sistema de RA es aquel que cumple con los siguientes 3 requerimientos [3]:

1. Combina la realidad y lo virtual. Al mundo real se le agregan objetos sintéticos que pueden ser visuales como texto u objetos 3D (wireframe o fotorealistas), auditivos, sensibles al tacto y/o al olfato.
2. Es interactivo en tiempo real. El usuario ve una escena real con objetos sintéticos agregados, que le que ayudarán a interactuar con su contexto [4].
3. Las imágenes son registradas en espacios 3D. La información virtual tiene que estar vinculada espacialmente al mundo real, de manera coherente. Se necesita saber en todo momento la posición del usuario, respecto al mundo real, y de esta manera, puede lograrse el registro de la mezcla entre información real y sintética.

Desde un punto de vista más amplio, la RA es una aplicación interactiva que combina la realidad con información sintética - tal como imágenes 3D, sonidos,

videos, textos, sensaciones táctiles – en tiempo real, y de acuerdo al punto de vista del usuario [5].

En general, los autores coinciden en que todo sistema de realidad aumentada ejecuta, de manera secuencial, las siguientes cuatro tareas (ver Figura 2):

1. Captura del escenario.
2. Identificación de la escena.
3. Mezclado de la realidad más aumento de información.
4. Visualización de escena aumentada.

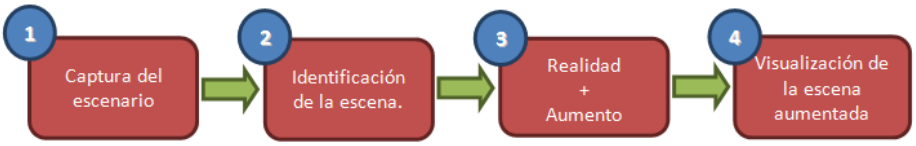


Fig. 2. Tareas de un sistema de realidad aumentada

2.1 Realidad Aumentada en el ámbito de educación y educación especial

La RA es una tecnología que está introduciéndose en nuevas áreas de aplicación, tales como: la reconstrucción del patrimonio histórico, marketing, diseño, arquitectura, entre otros. El mundo académico no está al margen de estas iniciativas, y también, ha empezado a introducir la tecnología de la RA en algunas disciplinas. Sin embargo, el conocimiento y el uso actual de RA en procesos educativo aún resultan novedosos y de poca frecuencia; entre otros motivos, se debe a la propia naturaleza y estado de desarrollo de dicha tecnología, así como también a su escasa presencia en los ámbitos cotidianos de la sociedad.

Una de las aplicaciones que han tenido buena difusión es la del proyecto Magic Book[1]. Se propone al alumno la lectura de un libro real a través de un visualizador de mano, y puede observar sobre las páginas reales contenidos virtuales. De esta manera, cuando el alumno ve una escena de RA puede introducirse dentro de la escena y experimentarla en un entorno virtual inmersivo.

Otro ejemplo de este tipo de aplicación es el "Curso para la mejora de la capacidad espacial" con RA, accesible desde AR-Books[2]. En éste, los estudiantes aprenden dibujo técnico, a partir de un libro en el que aparecen las proyecciones de objetos y, utilizando marcadores o patrones, pueden ver y mover los mismos objetos en 3D.

Uno de los aspectos más prometedores de la RA es que puede ser utilizada para favorecer varias formas interactivas de aprendizaje, unido a la gran facilidad con la que se superponen datos con el mundo real, permitiendo así que se simulen procesos dinámicos.

Otra característica clave de la RA para el ámbito educativo, es su capacidad para responder a las entradas del usuario. Esta interactividad le confiere un gran potencial

[1]http://www.hitlabnz.org/index.php/research/augmented-reality?view=project&task=show&id=54

[2]http://www.bubok.es/libros/202659/Curso-para-la-mejora-de-la-capacidad-espacial (Recuperado en 2013).

para el aprendizaje. Así, obliga a la participación de la persona, y se puede afirmar entonces que la RA es atractiva porque se alinea con el aprendizaje activo.

Por último, la RA basada en los dispositivos móviles aprovecha esta herramienta cada vez más omnipresente, no sólo para el desarrollo de las interacciones sociales, sino también para el aprendizaje y la investigación, desdibujando los límites entre el aprendizaje formal e informal, lo que a su vez permite contribuir a la evolución de una ecología de aprendizaje que trasciende las instituciones educativas [4][6].

Por otra parte, varios son los autores que remarcan el beneficio de la tecnología de realidad aumentada en el ámbito de educación especial. En este sentido, y en base a una revisión de antecedentes de aplicaciones de RA, en el área de discapacidad, los sistemas pueden clasificarse en [7]:

- Orientados a Personas con discapacidad visual
- Orientados a Personas con deficiencia auditiva
- Orientados al proceso de aprendizaje de personas con deficiencia intelectual
- Orientados a favorecer la interacción con la computadora

Se mencionarán aquí algunos ejemplos de aplicaciones, basadas en RA, que brindan ayuda a personas con algún tipo de limitación, ya sea física o mental, como una revisión de antecedentes específicos de RA en el ámbito de la educación especial.

PictogramRoom [8], es un proyecto que involucra una habitación de realidad aumentada para enseñar a comprender los pictogramas que permiten la comunicación a personas con trastornos del espectro del autismo, entre otros. El proyecto plantea que con la ayuda de la RA, se posibilita el uso de pictogramas superpuestos sobre objetos reales, y esto, es beneficioso para ayudar a visualizar la conexión entre imagen real y pictograma en tiempo real.

Eyering [9], es una creación del MIT Media Lab. Se trata de un anillo de realidad aumentada equipado con una pequeña cámara, un procesador, conectividad Bluetooth y retroalimentación auditiva, a través de un dispositivo portátil, que podría ayudar a las personas con dificultades visuales a identificar objetos y leer texto. Aunque también, podría funcionar como ayuda de navegación o traducción para cualquier persona, y en el entrenamiento para enseñar a leer a los niños. Un ejemplo de su utilización en el ámbito de la discapacidad, es cuando una persona no puede ver correctamente lo que tiene enfrente, entonces puede hacer uso del EyeRing, el cual le comunicará mediante un dispositivo móvil lo que la cámara ha captado, agregando información sonora a la escena capturada.

Otro proyecto es “e-labora” que incorpora la realidad aumentada y la tecnología 3G en actividades de entrenamiento y formación profesional. Este esfuerzo pretende mejorar la integración de las personas con discapacidad intelectual en el lugar de trabajo, creando un entorno que mejore su seguridad, su estabilidad emocional, su capacidad de comunicación, su autodeterminación y su participación. Por ejemplo, algunas aplicaciones incluyen una herramienta de navegación e información en el entorno de trabajo y una guía fácil sobre cómo utilizar equipos como una impresora [10].

También se han investigado otros proyectos de RA en el ámbito de la discapacidad tales como Babelfisk [11], Elcano [12], entre otros. Y otros más generales, en vinculación con RA y su aplicación en educación, [13], [14].

3 Motivación para diseñar AuthorAR

Si bien los proyectos estudiados presentan un particular interés, dada su aplicación al ámbito educativo y a personas con algún tipo de discapacidad en particular, ninguno de ellos está especialmente orientado a la creación de actividades educativas. Es sabido que existe, hoy en día, un gran número de herramientas de autor que facilitan la tarea de los docentes ofreciendo plantillas para la creación de diferentes tipos de actividades didácticas. Sólo a modo de ejemplo, se mencionan algunos programas que son considerados herramientas de autor orientados al ámbito educativo: jClic, Ardora, HotPotatoes, Lim, Malted, Adobe Director, ExeLearning, Constructor, Cuadernia, entre muchos otros. Varias de éstas son utilizadas por docentes y terapeutas para trabajar diferentes actividades con personas con necesidades especiales [15].

Sin embargo, son pocas las herramientas de autor, orientadas al ámbito educativo, que permiten generar actividades de realidad aumentada. Se revisaron algunas como: Cuadernia, Atomic y Aumentaty Author. Se detalla a continuación una breve descripción de cada una:

- Cuadernia[3]: es una aplicación para la creación de libros digitales con contenido multimedia que, a partir de su versión 2.0, permite añadir contenido de realidad aumentada. Para la creación de una actividad de RA, existe una plantilla específica en que se puede elegir un objeto del tipo .DAE[4] y especificar con qué rotación quiere visualizarse.
- Atomic Authoring Tool[5]: es un software de escritorio, multi-plataforma para la creación de aplicaciones de realidad aumentada. Es una capa de abstracción para la biblioteca ARToolKit.
- Aumentaty Author[6]: es una herramienta de autor, que permite la construcción de contenido RA, sin tener conocimiento en programación. El contenido se construye, a través de una interfaz gráfica de usuario. Utiliza tecnología de marcadores para reconocer el espacio tridimensional, mostrado por la webcam y posicionar el contenido. La herramienta se complementa con otra, llamada Aumentaty Viewer, necesaria para poder visualizar los proyectos de RA generados con Aumentaty Author.

Si bien, las tres herramientas permiten la generación de contenido de realidad aumentada, las únicas actividades posibles son las del tipo exploratoria. Es decir, a un marcador, se le puede asociar un objeto 3D que, dependiendo de la aplicación que se use, pueden ser de un tipo u otro.

Otra observación realizada, es que ninguna de estas herramientas dispone de un espacio para la inclusión del enunciado que le indique al alumno la intención de esta actividad, desde el punto de vista cognitivo y didáctico, ni tampoco la posibilidad de incluir algún tipo de retroalimentación (“feedback”) de audio y/o texto.

[3] http://cuadernia.educa.jccm.es/
[4] Es un esquema XML para la distribución e intercambio de recursos 3D entre aplicaciones.
[5] http://www.sologicolibre.org/projects/atomic/es/
[6] http://www.aumentaty.com/es/content/aumentaty-author

En la próxima sección se presenta el aporte de los autores, a partir del diseño de una herramienta de autor orientada a la creación de actividades educativas de RA, con la inclusión de plantillas específicas para el escenario de educación especial, con foco en el entrenamiento de competencias comunicacionales.

4 AuthorAR: Una herramienta de autor, basada en el paradigma de realidad aumentada.

4.1 Características Generales

En esta sección, se describe una herramienta de autor para la creación de actividades educativas, basadas en el paradigma de realidad aumentada, llamada AuthorAR.

La misma está pensada como una aplicación de escritorio, de libre distribución, que puede ser ejecutada en cualquier PC, y utiliza como método de reconocimiento de escenarios la tecnología de marcadores.

Se orienta específicamente al uso por parte de docentes y, en particular, se han diseñado algunas plantillas de actividades, cuyo propósito es el trabajo con personas del escenario de educación especial.

AuthorAR, cuenta con dos componentes, por un lado el generador de materiales educativos, y por otro lado, el visor.

El generador (orientada a docentes), está desarrollado en Java y permite crear fácilmente actividades educativas a partir de plantillas. Al momento, se han diseñado dos plantillas: una para actividades de exploración y otra para actividades de estructuración de frases. Para ambos tipos de actividades, la herramienta genera un .zip, que contiene un archivo de especificación en formato XML y una carpeta con los recursos a utilizar en la actividad. En la Figura 3, se puede ver la estructura del archivo comprimido generado por la aplicación y el formato del XML asociado.

Fig. 3. Estructura del archivo comprimido generado por AuthorAR

El visor (orientada al trabajo del alumno), es una herramienta desarrollada en ActionScript, que permite la visualización y resolución de las actividades educativas creadas con el componente generador de AuthorAR. Se mostrarán algunos ejemplos posteriormente.

4.2 Tipos de actividades disponibles al momento

La herramienta AuthorAR está en evolución, actualmente sólo hay disponibles dos tipos de plantillas, pero el proyecto abarcará una mayor variedad, orientadas a diferentes disciplinas, tipos de actividades (relación, ordenamiento, completamiento, entre otras), y niveles educativos. Los dos tipos de actividades ya desarrolladas, pueden vincularse (aunque no necesariamente) al área de educación especial.

4.2.1 Actividades de exploración

Las actividades de exploración son aquellas en las que el docente puede relacionar contenido multimedia a un marcador de RA, de manera tal que, los alumnos puedan utilizarlo como elemento de interactividad con el ordenador.

Generalmente, se puede calificar de actividad de exploración a toda aquella que provoca un nuevo aprendizaje, ya sea un nuevo concepto, una nueva regla, una nueva fórmula, nuevos saberes particulares. En general, se recurre a lo "simple" para ayudar a comprender lo complejo, y para ayudar a regresar, luego a esto [16].

En una actividad de exploración, el alumno es el actor principal, en la medida en que debería ser realizada por él mismo, y no por el docente.

Las actividades de exploración de AuthorAR, se basan en un conjunto de elementos que al ser presentados a la cámara web conectada a la pc, ofrecen información adicional del escenario que se está capturando en tiempo real. En primer lugar, durante la resolución de la actividad, aparece en la pantalla la consigna específica, elaborada e ingresada por el docente. Dicha consigna, puede aparecer en formato texto, audio o ambos. Luego, el alumno debe mostrar el marcador frente a la cámara con el objetivo de obtener información adicional, en forma de sonido y objetos 3D, en tiempo real. Un ejemplo de aplicación de esto, se vincula con actividades basadas en el reconocimiento de formas y colores. El entrenamiento de procesos de abstracción en alumnos usuarios de comunicación aumentativa y alternativa es muy importante, ya que permite luego reconocer en los pictogramas los objetos/acciones que la persona desea comunicar.

4.2.2 Actividades de estructuración de frases

Las plantillas de estructuración de frases permiten a los docentes generar actividades en las que el alumno deberá componer una frase del estilo sujeto-acción-objeto. Esto lo realiza a partir de la utilización de cartones como los que se muestran en la Figura 4, con imágenes que representan al sujeto, a la acción y al objeto, respectivamente. El docente puede pegar a la imagen su marcador correspondiente en el reverso o trabajarla tal como aparece en la Figura 4.

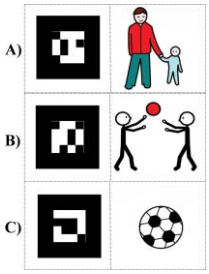


Fig. 4. Cartón para la interacción con la aplicación. A) Muestra a la persona B) Muestra la Acción C) Muestra el objeto

El alumno, podría aumentar la imagen asociada en cada marcador, presentándolo frente a la cámara web, similar a las actividades de exploración. En la Figura 5, se muestra cómo sería la interacción de un alumno con la aplicación, al mostrar el cartón de la pelota, incluido en la Figura 4 C). Como puede observarse, el objeto que se visualiza es 3D. Además, esta escena aumentada puede estar acompañada de un audio pertinente, si el docente así lo hubiese configurado, acorde a los objetivos de la actividad.

Fig. 5. Interacción con la aplicación con un único marcador. Visualización 3D del objeto.

Luego, para realizar el ejercicio de estructuración de frases, el alumno deberá presentar en forma conjunta y ordenada los cartones, de manera tal de componer la frase solicitada. Al presentarlos frente a la cámara web, conjuntamente y en el orden correcto, podrá visualizar al sujeto realizando la acción con el objeto correspondiente (Figura 6). Esto permitirá reforzar el mecanismo de estructuración de frases.

Fig. 6. Actividad de estructuración de frases.

Si la estructura de la frase es incorrecta, se podrán mostrar distintos tipos de retroalimentación o feedback, acorde a lo que el docente haya establecido en la herramienta de autor. Una aplicación de uso de la plantilla de estructuración de frases, está asociada al entrenamiento de alumnos con algún tipo de deficiencia en la adquisición del lenguaje. Los terapeutas y docentes suelen abordar el entrenamiento, a partir de la construcción de frases estructuradas. En este caso AuthorAR, buscaría potenciar la motivación y la estimulación de la persona, a partir de la interacción multimodal.

5 Propuesta de evaluación

Al momento se ha trabajado para el diseño de las plantillas con una fonoaudióloga que realiza el entrenamiento de prácticas para la adquisición del lenguaje, quien ha aportado sugerencias específicas acorde a su experiencia en este campo. En el corto plazo se realizará una evaluación con otros docentes abordando la construcción de actividades con estos dos tipos de plantillas disponibles en AuthorAR. Se pretende con esta evaluación, conocer la opinión de los docentes en relación a las posibilidades de creación de actividades, de manera tal de evolucionar esta herramienta, teniendo en cuenta la visión de los usuarios finales. Al mismo tiempo, se ha invitado a diferentes expertos en el área de educación, educación especial, y tecnología educativa a una sesión de focus group, de manera tal de someter a juicio de expertos el diseño actual, y analizar las próximas plantillas que sólo se tienen en diseño teórico.

En una segunda etapa, se procederá a evaluar un conjunto de actividades desarrolladas, con alumnos de escuelas e instituciones educativas, entre ellas algunas de educación especial con las que ya se tiene vinculación.

6 Conclusiones y trabajos futuros

Se cree que esta herramienta tiene fundamental importancia, ya que permite a los docentes construir actividades, basadas en realidad aumentada, establecer sus propias consignas y feedback, acorde a los objetivos didácticos planificados. Acorde al trabajo conjunto con una docente del área de educación especial, la plantilla vinculada a la estructuración de frases tendría un impacto positivo en el entrenamiento de procesos de comunicación, incorporación de vocabulario y de adquisición del lenguaje.

Si bien el proyecto está en desarrollo, ya se ha implementado una primera etapa que ha permitido plasmar parte del diseño en acciones concretas. Se está trabajando actualmente, en la primera evaluación de lo desarrollado al momento, y de los diseños por incorporar. Las plantillas ya construidas constituyen un aporte acorde a los relevamientos previos acerca de herramientas de autor educativas basadas en RA.

7 Referencias

[1] Lin Ch. and Chao J.-T. Augmented Reality-Based Assistive Technology for Handicapped Children. International Symposium on Computer, Communication, Control and Automation, 2010.

[2] Milgram, P., Takemura, H., Utsumi, A., y Kishino, F.. Augmented Reality: A class of displays on the reality-virtuality continuum. Telemanipulator and Telepresence Technologies, 2351, 282-292. 1994

[3] Azuma R. (1997). A Survey of Augmented Reality. In Presence: Teleoperators and Virtual Environments. 6, 4 August 1997, 355-385.

[4] Azuma R., Baillot Y., Behringer R., Feiner S.K., Julier S. J., MacIntyre B. (2001). Recent Advances in Augmented Reality. In IEEE Computer Graphics and Applications. Nov-Dec 2001, 34-47.

[5] Abdulmushli M. (2012). Análisis de sistemas de realidad aumentada y metodología para el desarrollo de aplicaciones educativas. http://eciencia.urjc.es/bitstream/10115/7805/1/1112-MIIM-TFM-MazenAbdulmushliAlsirhani.pdf. Recuperado en 2012.

[6] Basogain X., Olabe M., Espinosa K., Rouèche C. y Olabe J.C.(2007). Realidad Aumentada en la Educación: una tecnología emergente. Online Educa Madrid 2007: 7ª Conferencia Internacional de la Educación y la Formación basada en las Tecnologías. Online Educamadrid'2007 Proceedings, pp. 24-29.

[7] Ong S.K., Shen Y., Zhang J., and Nee A.Y.C. (2011). "Augmented Reality in Assistive Technology and Rehabilitation Engineering". ISBN 978-1-4614-0063-9, pages 603 - 630.

[8] Pictogram room.(2012) http://www.pictogramas.org/proom/init.do?method=initTab. Recuperado en 2012.

[9] Eyering.(2012). http://www.digitalavmagazine.com/es/2012/08/13/el-mit-crea-un-dispositivo-de-realidad-aumentada-para-ciegos-activado-por-voz/. Recuperado en 2012.

[10] Tomás de Andrés, Mari Satur Torre. "e-Labora: todos capaces en el trabajo.". http://www.qualcomm.com/media/documents/files/wireless-reach-case-study-spain-augmented-reality-spanish-.pdf. Recuperado en 2013.

[11] BabelFisk. (2010) http://www.gearlog.com/2010/09/speech-to-text glasses use aug.php. Recuperado en 2012.

[12] González C., Martínez M. A., Villanueva F.J., Vallejo D., López J. C. (2011). Sistema para la navegación en interiores mediante técnicas de Realidad Aumentada. Disponible en: https://arco.esi.uclm.es/public/papers/2011-Ei3-carlos.gonzalez.pdf. Recuperado en 2013.

[13] Rodriguez Lomuscio J.P. (2011). Realidad Aumentada para el aprendizaje de Ciencias en niños de Educación General Básica. Disponible en: http://www.tesis.uchile.cl/tesis/uchile/2011/cf-rodriguez_jl/pdfAmont/cf-rodriguez_jl.pdf. Recuperado en 2013.

[14] De Pedro Carracedo J. (2011). Realidad Aumentada: Un Nuevo Paradigma en la Educación Superior. CAFVIR 2011. Disponible en: http://www.redusoi.org/docs/LibroActasCAFVIR2011.pdf#page=300. Recuperado en 2013.

[15] Sacco A., Soto Pérez J. (2009). Software libre para las necesidades educativas especiales. Revista Comunicación y Pedagogía Nº235-236, especial año 2009.

[16] Roegiers X.. Pedagogía de la integración. Cap 7: Las implicaciones de la Pedagogía de la integración en los aprendizajes. San José, 2007. ISBN: 978-9968-818-36-0.

Uma plataforma de edição de aulas acessíveis para professores: transformando aula em diversidade

Cristiani de Oliveira Dias[a], Eliseo Berni Reategui[b],

Programa de Pós-Graduação em Informática na Educação– Universidade Federal do Rio Grande do Sul (UFRGS) Brasil

cristianideoliveiradias@gmail.com, liliana@cinted.ufrgs.br, eliseoraetegui@gmail.com

Resumo Planos de aula são ferramentas bastante úteis que podem ter diferentes propósitos, tais como: servir como guia em sala de aula, propor a utilização de determinados recursos, definir a abordagem pedagógica do professor, apontar o perfil dos estudantes para os quais se construiu o plano, registrar os objetivos da aula com relação aos alunos (Hunter, 2002). Alguns estudos mostram que a maior parte dos professores se preocupa em planejar suas aulas (Reategui, 2011), apesar de muitos ressaltarem a falta de tempo para realização destes planejamentos (Guimarães, 2009). Esta dificuldade pode estar relacionada ao estigma de que desenvolver um plano de aula, por vezes, pode parecer uma tarefa complexa. Facilitar e instigar a criação desses planos pelos professores é o objetivo desta pesquisa. Para isso, propõe-se uma plataforma de edição de planos de aula que permita ao professor interagir com uma comunidade de educadores, buscando fazer com que a criação e compartilhamento de materiais deem aos professores um sentimento de empoderamento. A plataforma proposta funciona como uma rede social educacional, permitindo edição de aulas e compartilhamento dessas entre participantes. Um sistema de recomendação busca localizar indivíduos com interesses similares para colocá-las em contato, com o objetivo justamente de instigar a socialização e colaboração entre estes. Este artigo apresenta uma revisão dos portais usados pelos professores brasileiros que servem para aprimorar as práticas em sala de aula. Apresenta também a descrição da plataforma em curso de desenvolvimento, a qual tem como principal característica a recomendação de pessoas para que possam interagir, trocar experiências, colaborar no desenvolvimento de seus planos de aula. Este artigo está organizado da seguinte forma: dar uma breve descrição do que são planos de aula, após na seção 2, onde descrevemos a plataforma de edição de aulas Educa, a seção 3 onde tratamos da construção de materiais para diversidade, logo depois na sessão 4 e subsessões nos referimos à plataforma Educa e suas características propostas no trabalho e por fim, a sessão apresenta considerações finais sobre a pesquisa realizada, e propõe direcionamentos para trabalhos futuros.

Palavras chave: Plataforma de edição de aulas, materiais educacionais, acessibilidade.

Abstract *Lesson plans are very useful tools that can have different purposes, such as serving as a guide in the classroom, proposing the use of certain resources, define the pedagogical approach of the teacher, pointing the profile of students for which the plan is built , record the lesson objectives with respect to students (Hunter, 2002). Some studies show that most teachers care in planning their lessons (Reátegui, 2011), although many emphasized the lack of time for completion of these plans (Guimarães, 2009). This difficulty may be related to the stigma that develop a lesson plan, sometimes it can seem like a complex task. Facilitate and instigate the creation of these plans by teachers is the goal of this research. For this, we propose a platform editing lesson plans that allow the teacher to interact with a community of educators, seeking to make the creation and sharing of materials deem teachers a sense of empowerment. The proposed platform works as an educational social network, allowing editing and sharing lessons among these participants. A recommendation system search locate individuals with similar interests to put them in touch with the purpose of instigating just socializing and collaboration between them. This article presents a review of the portals used by Brazilian teachers that serve to*

improve practices in the classroom. It also presents the description of the platform under development, which has as main feature the recommendation of people so that they can interact, share experiences, collaborate on developing their lesson plans. This article is organized as follows: give a brief description of what they are lesson plans, after the section 2, we describe the editing platform Education classes, section 3 where we treat building materials for diversity, right after the session 4 and subsections referring to the platform and Educa characteristics proposed in the paper, and finally, the session presents final considerations about the research, and proposes directions for future work.

Keywords: *Platform of Lesson Plans, educational materials, accessibility.*

INTRODUÇÃO

Vivemos atualmente em uma sociedade na qual se espera que todas as pessoas possam participar dos diferentes espaços sociais. Ao mesmo tempo, essa sociedade preconiza uma educação inclusiva. Educação inclusiva refere-se ao processo educativo embasado no paradigma de inclusão, segundo o qual toda pessoa deveria ser capaz de ter oportunidade de escolha e de autodeterminação (MITTLER, 2003). Para o autor, uma educação inclusiva não implica em colocar todas as crianças em escolas, mas em transformar as escolas para torná-las mais responsivas às necessidades dos seus alunos, e em ajudar seus professores a aceitar a responsabilidade pela aprendizagem dos seus alunos. Como desafio preconizado pela Web 2.0, existe um novo perfil de usuário. De meros receptores de informações, tornam-se criadores, desenvolvedores de conteúdos podendo, com isso, criar, produzir materiais e depois compartilhar com colegas e professores na Internet. Atualmente, um dos desafios do professor é acompanhar os alunos no mundo digital e ir além disso: usar estes recursos para tornar suas aulas mais próximas da forma como os alunos se motivam a aprender. Entendemos, portanto, que os professores poderão fazer parte dessa interação com o aluno, desenvolvendo materiais educacionais digitais mais ricos e que contemplem a todos os alunos. A diversidade midiática pode favorecer a inclusão desses alunos com necessidades educativas especiais, assim como a adaptação desses materiais educacionais digitais. Este artigo tem como objetivo principal propor uma plataforma de edição de planos de aula, a fim de contemplar a diversidade de alunos em sala de aula e principalmente recomendar requisitos para a construção de planos acessíveis a pessoas com deficiência.

O interesse nesta pesquisa se deu a partir de estudos dos autores na área de educação especial, bem como na participação em projeto de pesquisa de construção de plataforma de edição de aulas com recomendação de conteúdos (Acosta et al, 2012).

O que são Planos de Aula?

Os planos de aula correspondem a uma proposta de organização do processo de ensino e aprendizagem. Segundo Nikolic e Cabaj (2000) nenhum plano de aula funciona bem para todos os grupos de alunos. Podem assumir formas diferentes e incluir uma ampla variedade de conteúdos. Para isso, o professor precisa criar o plano tendo como base a estrutura da sua classe e a variedade de conhecimento dos seus alunos, atendendo os objetivos do conteúdo e a multiculturalidade dos estudantes. O plano também deve prever a possibilidade de que esse conteúdo e práticas propostos atinjam o objetivo determinado. A figura 1 exemplifica uma situação em que um plano de aula é aplicado sem objetivos definidos. O resultado disso é uma aula que não envolve os alunos, não trabalha objetivamente nenhum processo de aprendizagem específico. Os pontos de interrogação representam os alunos e as setas pontilhadas representam o plano de aula criado pelo professor. Quando um plano de aula é bem desenvolvido, presume-se que seu propósito possa atingir diretamente os alunos, o conteúdo proposto e as características de cada sala de aula.

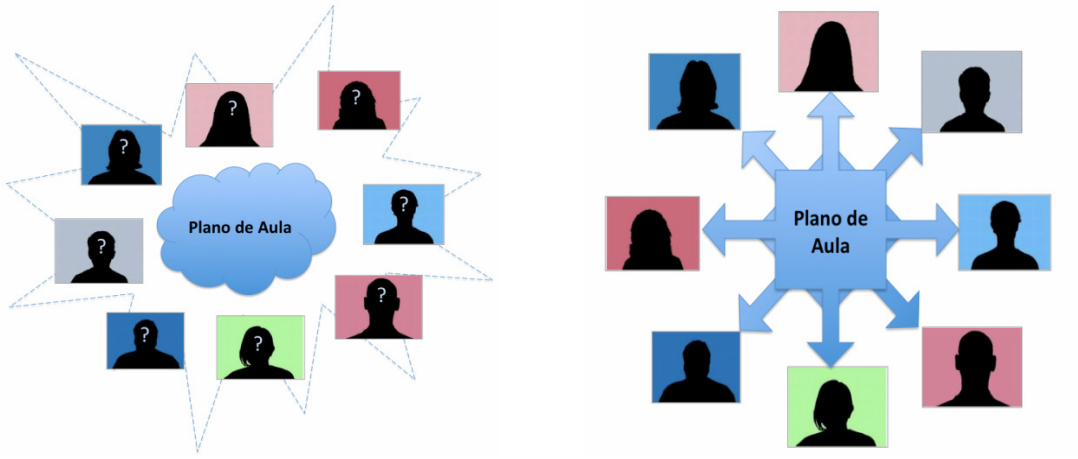


FIGURA 1. Modelos de planos de aula, adaptados de nikolic e cabaj (2000)

Tão logo o professor tenha clareza sobre as questões relacionadas ao que ensinar, surge a questão: como trabalhar? Em geral, a sequência de aula deveria seguir a sequência natural do processo de aprendizagem, procurando respeitar os conhecimentos prévios dos estudantes. Frequentemente, num primeiro momento são apresentados novos conteúdos. Num segundo momento, os alunos têm a oportunidade de trabalhar de forma mais ativa, experimentando, cometendo erros e recebendo feedback, sendo corrigidos e tentando de novo (Nikolic e Cabaj, 2000).

A elaboração de planos de aula envolve objetivos, conteúdos, procedimentos, recursos de ensino e avaliação. Na prática cotidiana de nossas escolas, o planejamento tem constituído uma tarefa burocrática e sem sentido, representando, muitas vezes, apenas um documento a mais para o arquivo do coordenador. É a separação entre o pensar e o fazer, segundo a lógica do sistema (Martins,1989).

Neste sentido, é importante ressaltar que o planejamento não é um fim em si mesmo, mas um meio de preparar e organizar a ação tendo em visto um objetivo.

Material para Diversidade – A inclusão em sala de aula

Para auxiliar esse processo de construção de uma aula atendendo a diversidade e que facilite essa criação pelo docente, os planos de aula são importantes informações de como serão desenvolvidas as aulas, apresentação de conteúdo, criação de atividades, interação e colaboração entre pares,pensa-se que a criação de materiais para diversidade são de extrema importância.

Segundo os dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) as matrículas na educação especial estão distribuídos entre as redes de ensino municipal (38,2%), privada (37,3%) e estadual (24,3%) sendo que entre as escolas públicas, predomina o atendimento educacional em classes comuns do ensino regular, pois 69,7% das matrículas da educação especial, o que evidencia a anuência dos gestores e educadores públicos à proposta de fortalecimento da inclusão educacional[1]. Partindo desta realidade torna-se imperativo não apenas pensar a inclusão no contexto educacional, mas a permanência e acesso a todos os espaços por parte desses alunos com necessidades educacionais especiais. O professor, ator importante nesse processo inclusivo, é responsável pelo planejamento, seleção e construção de materiais didáticos que servirão de base para o processo de construção de conhecimento.

Santos (2007) reafirma que

> Para a ação docente no contexto da diversidade, necessário se faz trabalhar com redes de encontros. Encontro de saberes, fazeres, reflexões, metodologias, estratégias de ensino, recursos, perspectivas avaliativas (...)

1 O Programa Educação Inclusiva: Direito à Diversidade, vinculado à Secretaria de Educação Especial do Ministério da Educação (2003), objetiva fomentar a política de construção de sistemas educacionais inclusivos nos municípios brasileiros.

Esse papel mediador do professor (mediar os processos de criação de materiais para satisfazer o perfil de seu aluno) é muito importante, pois nele são projetados os seus conceitos epistemológicos como sugere Zaballa (1998, p.33) onde fala que "entre as diversas correntes existentes (...) há uma série de princípios nos quais as diferentes correntes estão de acordo: as aprendizagens dependem das características singulares de cada um dos aprendizes." O mesmo autor enfatiza a importância do educador "observar a atenção a diversidade dos alunos com eixo estruturador" (p.34) e o mesmo autor alerta para o fato de esquecermos em alguns momentos desses processos de aprendizagem. Mas como disponibilizar esses materiais educacionais para que pessoas com necessidades especiais tenham a eles acesso?

A investigação acerca dos materiais educacionais acessíveis para apoiar o processo inclusivo iniciou em 2008 (Dias). A problemática partiu na pesquisa de como os materiais educacionais poderiam atender à diversidade. Para nós, materiais educacionais que atendem à diversidade significam que sejam construídos desde seu projeto pensando em atender a pessoas com necessidades educativas especiais (PNEEs). Essa preocupação não é recente, diversas pesquisas abordam a questão da acessibilidade e da importância de existir materiais adaptados às diferentes necessidades dos alunos (Poletto, 2007). Porém, na maioria das vezes a proposta é re-adaptar o material para uma determinada necessidade ou limitação. Como por exemplo, os materiais educacionais desenvolvidos pelo Cefet-BG[2]. Em suma, os materiais que se dizem acessíveis, são na verdade "novas" versões diferenciadas dos materiais educacionais originais e que de alguma forma excluem esse aluno com necessidades especiais, ao não permitir a interação no mesmo tipo de material, privando o aluno ao acesso à informação original em sua completude. Obviamente não se trata de manter a versão original do material de forma que não permita a interação desses alunos, mas pensar tal material de forma **convergente midiaticamente** falando.

Esta visão é o cerne de um verdadeiro processo inclusivo, trazendo as pessoas para desenvolverem suas ações em conjunto com outras sem diferenciações. Nesse sentido, Warschauer (2006) compartilha do mesmo pensamento conceituando a inclusão social como "não apenas uma questão referente à partilha adequada de recursos, mas também de participação na determinação das oportunidades de vida tanto individuais quanto coletivas (p.25)". Com a popularização da Internet e a construção de espaços de cidadania virtuais percebe-se a necessidade de prover esse mesmo acesso universal agora no ciberespaço e inicia-se então a preocupação com a inclusão digital. A inclusão digital e social são fortemente imbricadas como afirma Warschauer (2006) "a capacidade de acessar, adaptar e criar novo conhecimento por meio do uso das novas TICs[3] é decisivo para a inclusão social na época atual (p. 25)". Vale ressaltar que a utilização de máquinas e conectividade não viabiliza por si só a inclusão digital e sim o processo de práticas sociais apoiados na realidade de cada comunidade com a utilização dos recursos das TICs, favorecendo assim a autonomia e inclusão dos sujeitos. Frente a isso, Passerino (2005) afirma que a utilização das TIC pode promover o desenvolvimento social, afetivo e cognitivo de todos os sujeitos, particularmente, o das PNEEs[4]. No entanto, questiona-se até que ponto os materiais educacionais desenvolvidos estão atentos a essa parcela da população. Por outro lado, é importante estabelecer quais são as necessidades sentidas por este público ao utilizarem esses recursos que tanto podem contribuir para sua autonomia social, cultural e cidadã. Na medida em que se pretende investigar a inclusão social de PNEEs via materiais educacionais, vê-se uma ligação direta com a acessibilidade destes materiais educacionais.

De acordo com esse pensamento, concordamos com os autores quando "o sucesso não pode ser definido de forma linear, tendo em consideração produtos iguais para todos. A diversidade de competências dos alunos terá que corresponder à diversidade de produtos a considerar em termos de sucesso" (Amaral e Ladeira, 1999). Pesquisando sobre inclusão social e digital, encontra-se um conceito chave que é o de acessibilidade, sendo ela condição necessária à inclusão digital. O conceito de acessibilidade surge ligado a questões físicas relativas a facilidades de acesso (barreira arquitetônicas) e reabilitação física e profissional. Posteriormente é transferido para a informática na questão de acesso à Web especificamente e transforma-se em metas de desenvolvimento para todos os países[5].

---

[2] Materiais criados e publicados no endereço: http://www.bento.ifrs.edu.br/ept/oa/regradetres/#iniciomenu

[3] Significado de Tecnologias de Informação e Comunicação.

[4] Nesse artigo escolhemos utilizar a terminologia Pessoas com Necessidades Educativas Especiais.

5 O conceito de acessibilidade surge no Brasil já fazendo referência também aos meios de comunicação. Essa amplitude do termo está contemplada no Decreto Lei nº 3.298 de 1999 que definiu a acessibilidade na Administração Pública Federal como "possibilidade e condição de alcance para utilização com segurança e autonomia dos espaços, mobiliário e equipamentos urbanos das instalações e equipamentos esportivos, das edificações, dos transportes e dos sistemas e meios de comunicação" (BRASIL, 1999).

> "representa para o nosso usuário não só o direito de acessar a rede de informações, mas também o direito de eliminação de barreiras arquitetônicas, de disponibilidade de comunicação, de acesso físico, de equipamentos e programas adequados, de conteúdo e apresentação da informação em formatos alternativos" (Acessibilidade Brasil6, 2006).

Para Granollers (2004) acessibilidade significa proporcionar flexibilidade para adaptação às necessidades de cada usuário e a suas preferências e ou limitações. Partindo dessa ideia, este artigo propõe uma plataforma para construção de planos de aula com recomendação de materiais acessíveis, apresentando também dicas para a criação destes materiais.

Plataforma EDUCA, Sistemas de Recomendação e Acessibilidade

Acreditamos que o ponto de convergência entre esse processo de construção didática e pedagógica traz a tecnologia como aliada. Esse processo de construção de materiais didáticos, de pensar no conteúdo contemplando toda a sala de aula é uma função dificultosa para o docente. Utilizar uma ferramenta/plataforma como auxílio, um braço que ajude nessa construção nos leva a acreditar que esse projeto é de grande valia para educação inclusiva.

A ferramenta desenvolvida neste trabalho foi chamada de EDUCA, e comporta-se como uma plataforma onde o professor cadastrado pode criar e editar suas aulas. A figura 3 mostra a tela de edição de aulas da plataforma, na qual está sendo editada uma aula sobre "Professor e inclusão em sala de aula".

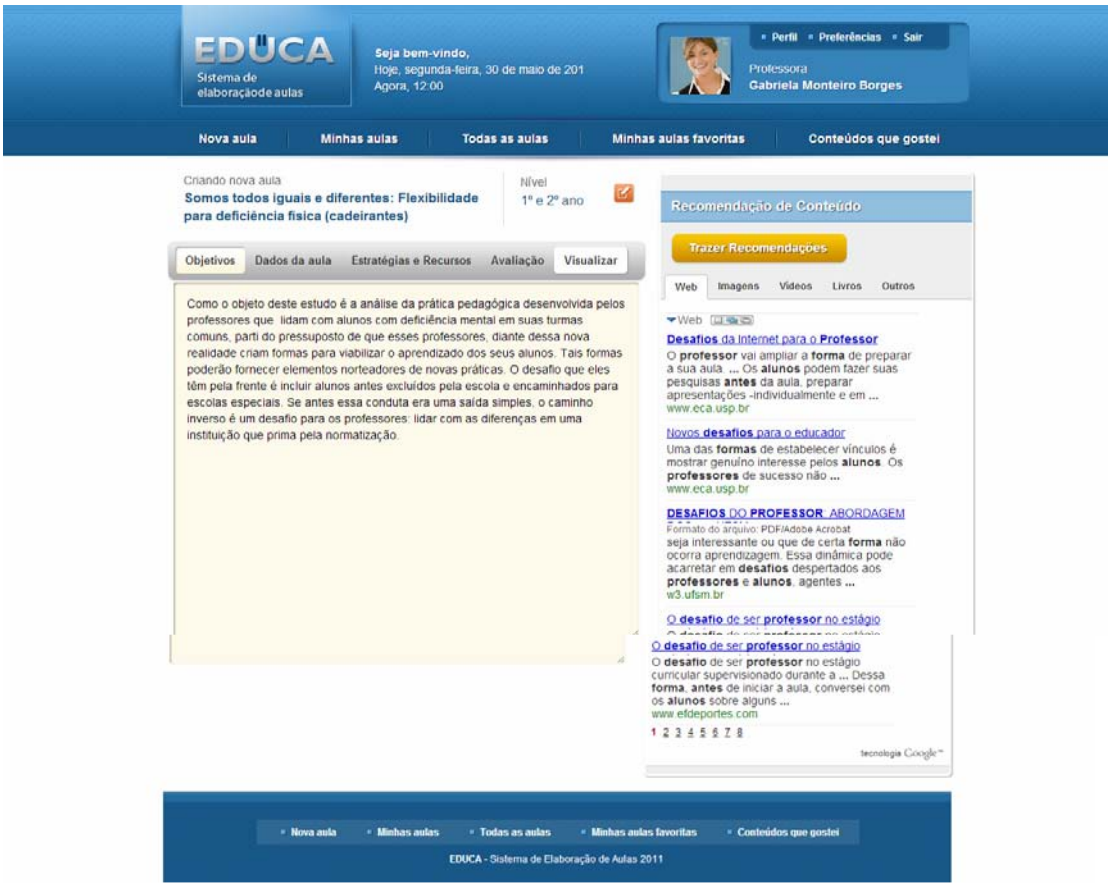


FIGURA 3. Plataforma Educa

---

[6] Disponível em http://www.acessobrasil.com.br

No lado esquerdo da tela, vê-se o texto sendo introduzido pelo professor sobre o tema da aula, no caso “Professor e inclusão em sala de aula”. No lado direito, observam-se duas áreas com funções específicas que buscam apoiar os professores na construção de seus planos. A primeira dela é um sistema de recomendação que busca na web informações referente às palavras-chaves identificadas pelo sistema por meio de uma ferramenta de mineração de texto (Reategui et al., 2011). Os conteúdos recomendados são páginas web, vídeos, livros, imagens e outros. A busca a estes conteúdos em tempo real é feita por meio de uma API do Google7 incluída na plataforma para otimizar a criação desses planos de aula. A figura 4 mostra alguns materiais recomendados automaticamente pela plataforma através da análise do conteúdo dos textos introduzidos até então pelo professor para sua aula.

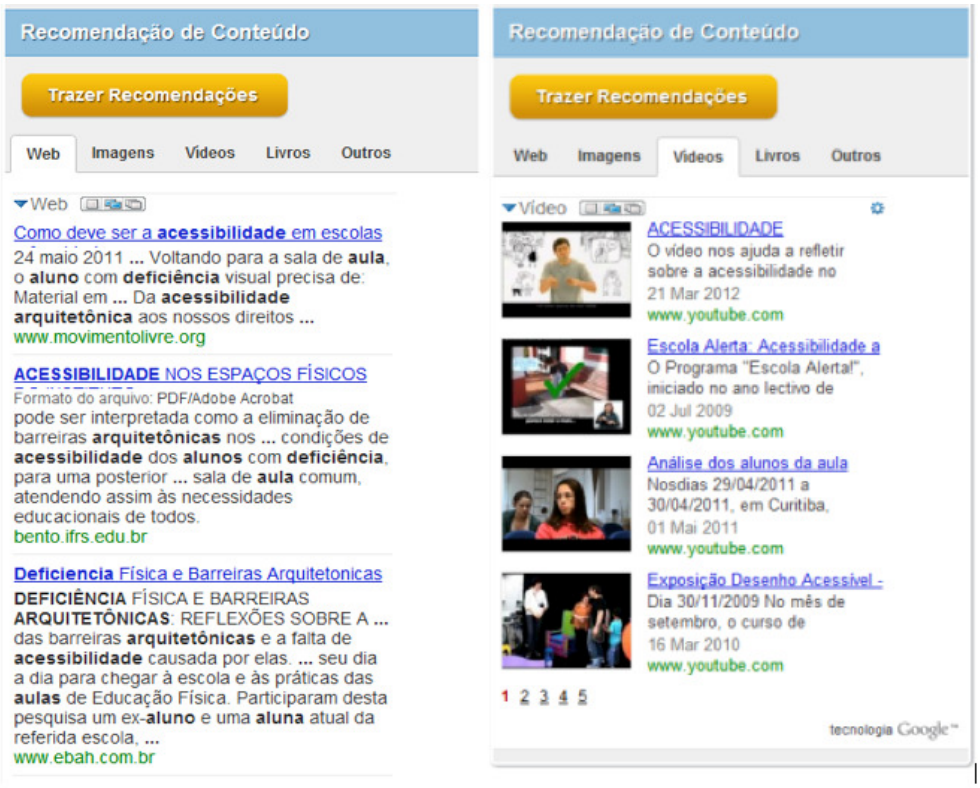


FIGURA 4. Recomendações de Textos (Lado Esquerdo Da Figura) e Recomendações de Vídeos (Lado Direito)

Os esforços direcionados para o desenvolvimento desta plataforma focam na criação de um sistema de edição de planos de aula fundamentados nos conceitos de "aprendizagem" e "comunidade", em vez de buscar fundamentação em modelos instrucionais rígidos (Barab e Duffy, 2000). Nossa perspectiva é consistente com a teoria e prática atual no desenvolvimento profissional docente, indicando que a mudança é mais provável que seja eficaz e duradoura se os professores estiverem autorizados a construir relações vitais entre si (Maverech, 1995; Richardson, 1990).

A interação é ponto principal desta pesquisa, pois com ela os professores poderão melhorar suas aulas, rever conteúdos antigos com modo de ensinar novos, conhecer a realidade de outras escolas além de criar materiais educacionais voltados às necessidades dos seus alunos com ou sem deficiência. As identidades coletivas são um processo permanente, dialógico, de pertencimento e partilha, de constituição de significações que orientam ações, como destacado por Castells (2000).

4.1. Da Implementação do Educa

A Plataforma Educa permite uma configuração específica de cada professor no momento do seu acesso, com isso o professor define quais os temas a serem abordados, quantos alunos o plano de aula irá atender e quais as deficiências desses alunos, se houver alguma.

[7] https://developers.google.com/custom-search/v1/overview

Para isso, foram propostos alguns questionamentos acerca do perfil do aluno, qual o tipo d deficiência, etc. no momento da criação de uma aula Dentro dessa perspectiva de criação de materiais de aula buscamos a metodologia de Design Centrado ao Usuário e alguns princípios são definidas como: **Desenho para os usuários** e suas tarefas que significa que o material educacional deve ser projetado para atender a todos os perfis de alunos, com opções de som (alunos cegos), vídeos com legendas (para alunos surdos), animações em tempo de acordo com a necessidade do aluno (para alunos com dificuldade motora) e se possível incluir também tecnologias assistivas[8]. **Consistência** significa manter os padrões utilizados na criação dos materiais. É importante que os formatos de apresentação de informações, estilos de fontes, cores, sejam usados de forma consistente e padronizada em todo o material, pois facilita a navegação e utilização pelos alunos. Como exemplo: em seu material está presente um botão chamado "voltar" de cor verde. O aluno irá relacionar a cor com a função, portanto toda vez que ele ver um botão verde, saberá que corresponde à função voltar. **Diálogo simples e natural** é importante para que o aluno entenda realmente o que está sendo pedido, palavras tecnológicas ou formais demais não levarão o aluno a uma resolução do problema mais eficaz. **Redução do esforço mental** do usuário e diminuição da carga cognitiva significa que quando o material é mostrado de forma que apresente som, vídeos e animações, todas no mesmo material e mostradas ao mesmo tempo, poderá levar o aluno à dificuldade em entender o que realmente o professor está pedindo como atividade. Material lúdico é sempre importante para prender a atenção do aluno, mas sobrecarregar de informações nem sempre será eficiente. **Proporcionar mecanismos adequados de navegação** é uma opção para mostrar ao aluno onde o aluno está situado no material educacional. Os alunos têm a tendência a "clicar" em todos os botões para ver o que o material irá fazer então mostrar ao aluno, em que atividade ele está e para onde ele vai é sempre importante para que ele consiga entender a sua proposta. **Deixar que os usuários dirija a navegação** serve para que o professor defina qual o tipo de estrutura de navegação que irá utilizar como: navegação linear[9], navegação hierárquica[10,] navegação não-linear[11] e navegação composta[12] (Amante e Morgado, 2001).

A plataforma armazena no banco de dados às informações que serão importantes no resultado da recomendação além de mostrar indicações no preenchimento da aula.

Para disponibilizar as informações de validação e verificação de acessibilidade, foi implementado um espaço de edição de textos13. Esse editor permite que o professor crie sua aula utilizando recursos como textos, vídeos, imagens, áudios e outros. O editor é caracterizado como uma ferramenta de criação que possibilita aos usuários produzir conteúdo em conformidade com as regras de acessibilidade.

Enfatizando o processo de validação de acessibilidade de conteúdo web, o professor ao criar seu material na Plataforma, tem opções diferentes para revisão da acessibilidade dos conteúdos do material produzido.

Para validar a acessibilidade do material desenvolvido foi utilizado o plug-in *AChecker* instalado dentro do editor de texto escolhido. Com esta opção o professor pode saber em que nível de conformidade está o material, tendo a opção de visualizar detalhes técnicos associados com a revisão automática como mostrado na figura 5.

Após a construção do material didático, o professor poderá navegar em páginas, verificar o conteúdo das páginas *web* adicionar esse material pesquisado no material que que está sendo criado, sendo que a própria Plataforma irá recomendar somente materiais que apresentam critérios em conformidade com as diretrizes da W3C (padrões de acessibilidade). Portanto, recomendar sites acessíveis também faz parte do estudo aqui definido.

---

[8] Tecnologias Assistivas que são conceituadas por Bersch (2009) como **Recursos** e **Serviços** que contribuem para proporcionar ou ampliar habilidades funcionais de pessoas com deficiência e conseqüentemente promover **Vida Independente** e **Inclusão**.

[9] Onde o aluno tem uma navegação seqüencial.

10 Essa navegação segue a lógica de especificação dos conteúdos a partir de conteúdo central.

11 A navegação é livre, o sistema não estabelece qualquer hierarquia ou seqüência de consulta aos conteúdos.

[12] Navegação livre, mas que ocasionalmente pode sugerir percursos lineares ou hierárquicos.

[13] Programa escolhido foi o TinyMCE, gratuito e *opensource.*

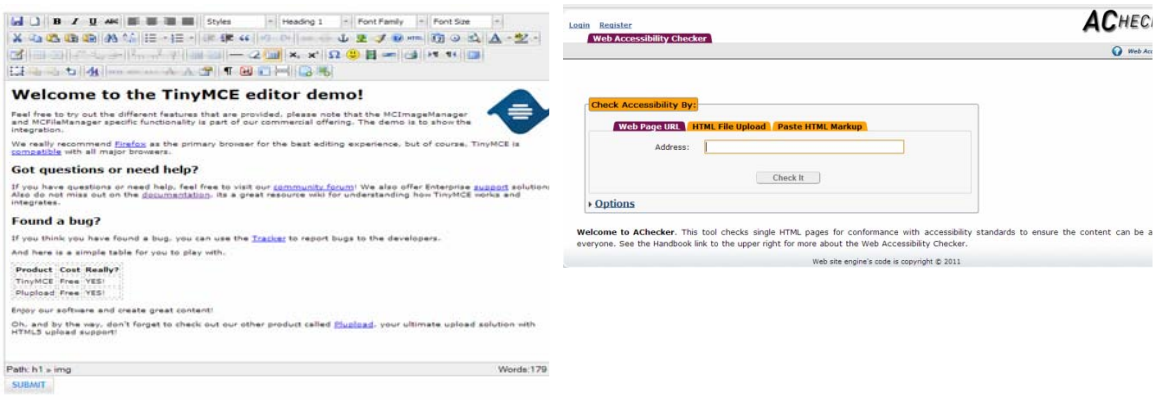


FIGURA 5. Tela do Editor Tinymce e Validador de Acessibilidade *Achecker*

Da Acessibilidade

Após digitar o material educativo no editor de texto e utilizar o recurso de recomendação, o professor pode utilizar o recurso de acessibilidade para verificar se o material produzido está acessível. Com isso, o sistema conta com uma API chamada AChecker que opera como um sistema de Web Service inserido no código fonte do editor de texto como mostra figura 6.

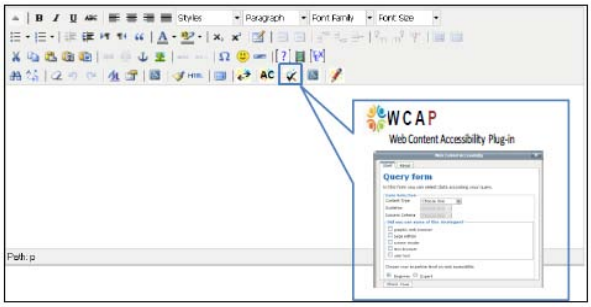


FIGURA 6. Editor já habilitado com o ícone de verificação de Acessibilidade

Com esta opção o professor pode saber em que parte do seu material educacional está ocorrendo erro em relação à acessibilidade. Este relatório identifica os elementos: uma ligação com a descrição detalhada dos resultados do validador automático, qual tipo de erro foi identificado e quais as recomendações dos especialistas (no caso das especificações da lista de camadas de orientação (WCAG/WAI). Alguns exemplos gerais de regras são definidas na tabela a seguir:

TABELA 1: *Web Content Accessibility Guidelines* 2.0 (WCAG),142008. (Construção da autora)

Princípio 1: Perceptível			
A informação e os componentes da interface do usuário têm de ser apresentados aos usuários em formas que eles possam perceber			
Alternativas em Texto	**Mídias com base no tempo**	**Adaptável**	**Discernível**

[14] http://www.w3.org/TR/WCAG20/

Isso significa que se o material educacional que está sendo construído tem muitas imagens ou animações, que sejam fornecidas alternativas em texto para qualquer conteúdo não textual.	Fornecer alternativas para mídias com base no tempo. Isso significa que sejam disponibilizadas mídias alternativas dentro do conteúdo apresentado. Que possa ser um vídeo, áudio e que apresente também legendas no caso de vídeo e autodescrição no caso das duas mídias.	Criar conteúdos que possam ser apresentados de diferentes maneiras (por ex., um layout mais simples) sem perder informação ou estrutura.	Facilitar a audição e a visualização de conteúdos aos usuários, incluindo a separação do primeiro plano e do plano de fundo.

Por exemplo, pode ser apontado que o material contém muito texto para aluno sem opção de recurso de leitor de tela (para cegos) ou linguagem inadequada sem tradução em libras (para surdos).

O desenvolvimento dessa plataforma está sendo pensada de forma a atender as necessidades do professor na construção do seu material e plano de aula. A tecnologia trabalha como um auxílio na produção de conteúdo, atividades, interação que se possa fazer para que o professor com isso consiga atender à diversidade em sua sala de aula.

CONSIDERAÇÕES FINAIS

O propósito desse artigo foi descrever uma ferramenta que possibilita ao professor a criação de planos de aula utilizando materiais educacionais que atendam à diversidade da sua sala de aula. Por meio desta ferramenta, busca-se permitir a utilização de várias mídias, atendendo assim o maior de número de alunos e suas necessidades. Auxiliar o professor a inserir várias mídias (vídeo, áudio, texto,etc) atendendo a necessidade de cada aluno é uma das propostas dessa plataforma também, denominando isso como convergência de mídias. Porém, que isso seja previsto na construção dos materiais educacionais digitais, não confundindo com a apresentação de todos ao mesmo tempo no mesmo material, evitando assim uma sobrecarga cognitiva.

Todas essas descrições apresentadas no corpo do texto como regras de acessibilidade, convergência de mídias fazem parte do relatório apresentado para o professor na hora da criação de seu plano de aula, elaborado por especialistas (definição criada pelos autores). Com isso, a recomendação oferecida possibilita um material diferenciado e que atenda a todos os tipos de alunos e modos de aprendizado. A Plataforma Educa está sendo implementada para também a oferecer a opção de agente no qual grava as configurações do professor, no momento de entrada na plataforma e "aprende" sobre os materiais educacionais construídos, tema, conteúdo e mídias já incluídas anteriormente, e critérios de acessibilidade recomendados afim de sugerir diferentes materiais em cada acesso, para cada aula, professor e necessidade.

REFERÊNCIAS

ACESSIBILIDADE BRASIL – O que é acessibilidade. Disponível em http://www.acessobrasil.org.br Acesso em 20 jul 2008.

BARAB, S.A., & Duffy, T. (2000). From practice fields to communities of practice. In D. Jonassen, & S.M. Land. (Eds.), Theoretical foundations of learning environments (pp. 25–56). Mahwah, NJ: Lawrence Erlbaum Associates.

BRASIL. Presidência da República. Casa. Decreto Lei 3.298 Disponível em https://www.planalto.gov.br/ccivil/decreto/d3298.htm. Acesso em 4 dez. 1999.

BRASIL. Decreto Nº 9.394, de 19 de dezembro de 2005. Disponível em: http://www.presidencia.gov.br/CCIVIL/_Ato2004/2006/2005/Decreto/D5622.htm. Acesso em: 10 jan. 2006.

BERSCH, Rita. Endereço: http://www.assistiva.com.br/. Acesso em 10 jul 2011.

CASTELLS, Manuel . A Era da Informação - Economia, Sociedade e Cultura, vol. 2: O Poder da Identidade. 2 ed. São Paulo: Paz e Terra, 2000.

FILATRO, A. Design Instrucional Contextualizado: educação e tecnologia. Pagina: 32. Editora: Senac. 2004.

GUIA – Grupo Português pelas iniciativas de Acessibilidade. [online] Disponível em URL:http://www.acessibilidade.net Acesso em 31 de maio de 2000.

Guimarães, A. (2009) O planejamento deve ser flexível. Revista Nova Escola, Janeiro.

GRANOLLLERS, T. MPIu Uma metodologia que integra la ingenieria del software, la interacción persona-ordenador y la accesibilidad en el contexto de equipos de desarrollo multidisciplinares. Tesis de doctorado, Universidad de Lleida, julio 2004.

Hunter, B. (2002). Learning in the Virtual Community depends upon Changes in Local Communities. In K. a. Renninger & W. Shumar (eds.), Building Virtual Communities. Learning and Change in Cyberspace.New York: Cambridge University Press, pp. 96□126.

LADEIRA, F.; AMARAL, I. A educação de alunos com multideficiência nas Escolas de Ensino Regular. Coleção Apoios Educativos. Lisboa: Ministério da Educação. Departamento da Educação Básica, 1999.

LEWIS, C. (2002). Lesson study: A handbook of teacher-led instructional improvement. Philadelphia: Research for Better Schools, Inc.

MARTINS, Pura Lúcia Oliver. Didática teórica, didática prática: para além do confronto. Edições Loyola, São Paulo, 1989.

MEVARECH, Z.R. (1995). Teacher's paths on the way to and from the professional development forum. In T.R. Guskey & M. Huberman (Eds.), Professional development in education: New paradigms and practices (pp. 151–170). New York: Teachers College Press.

MITTLER, M. Educação Inclusiva. Porto Alegre: ArtMed, 2003.

MONTARDO, S. P.; PASSERINO, L. Inclusão social via acessibilidade digital: proposta de inclusão digital para PNE. E-Compós, v. 8, p. 1-18, 2007. Learning Objects: Theory, Práxis, Issues, and Trends – Alex Koohang & Keith Harman, (eds), 2007. Santa Rosa, California, Informing Science Press.

NIKOLIC, Vesna; Caba, Hanna. 2000. Am I Teaching Well? Self – evaluation strategies for effective teachers. Pippin Publishing Corporation, Ontario Canada.

PASSERINO, L. Pessoas com autismo em ambientes digitais de aprendizagem : estudo dos processos de interação social e mediação. Tese de Doutorado. Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Informática na Educação 2005.

REATEGUI, E., Epstein, D., Lorenzatti, A., Klemann, M. (2011) Sobek: a Text Mining Tool for Educational Applications. In Proceedings International Conference on Data Mining (DMIN), Las Vegas. p. 59-64.

SONZA, A. P. (2008) Ambientes Virtuais Acessíveis sob a perspectiva de Usuário com Limitação. Tese (Doutorado). Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Educação, Porto Alegre, 2008.

WARSCHAUER, Mark. Tecnologia e Inclusão Social. A exclusão digital em
debate. São Paulo: Senac, 2006.

W3C/WAI. Web Accessibility Initiative Disponível em: http://www.w3.org/WAI. Acesso em: 06 abr. 2012.

WCAG/WAI. Web Content Accessibility Guideline Disponível em: http://www.w3.org/TR/WCAG20/. Acesso em: 06 abr. 2012.

ZABALLA, Antoni. A prática educativa: como ensinar. Artmed. Porto Alegre, 1998.

Aprendizaje Basado en Competencias y Objetos de Aprendizaje

Silvina Bramati, Zulema Beatriz Rosanigo, Claudia López de Munain, Pedro Bramati

Facultad de Ingeniería – Sede Trelew, Universidad Nacional de la Patagonia San Juan Bosco
Belgrano y Roca – Trelew – Chubut – Argentina
silvina.bramati@gmail.com, brosanigo@yahoo.com.ar, klaucvj@hotmail.com,
pedrobramati@speedy.com.ar

Resumen. En la actualidad hay consenso generalizado en desarrollar perfiles profesionales en base a competencias y es una tendencia internacional en el nuevo diseño de los planes de estudio el uso de las competencias como horizonte formativo, ocupándose no sólo de los saberes sino también del saber hacer y el saber ser. Facilitar el desarrollo de competencias durante el proceso de formación requiere un enfoque de la educación centrado primordialmente en el estudiante y en su capacidad de aprender, exigiendo mayor protagonismo y compromiso por parte del alumno. En este contexto, el diseño de materiales educativos orientados al desarrollo de competencias, cobra hoy en día una especial importancia. Los Objetos de Aprendizaje (OA) son una excelente alternativa para ello, permitiendo aprendizaje abierto al desarrollo de saberes (saber, saber hacer, saber ser). En este artículo se presenta una propuesta metodológica para diseñar OA como recursos orientados al desarrollo de competencias.

Palabras claves: competencias, objeto de aprendizaje, aprendizaje centrado en el estudiante.

Abstract. There is now widespread agreement on developing professional profiles based on skills and is an international trend in the design of new curricula. Facilitate the development of skills during training requires education approach focused primarily on the student and their ability to learn, requiring greater role and commitment of the student. In this context, design of educational materials aimed at developing skills, is of particular importance. Learning Objects (LO) are an excellent alternative, allowing open learning for the development of knowledge (knowing, knowing how, and knowing how to be). In this paper we present a methodology for designing LO as resources for skills development.

Keywords: skills, learning object, student-centered learning.

1 Introducción

El Aprendizaje Basado en Competencias (ABC) es un enfoque de enseñanza-aprendizaje que parte del conjunto de conocimientos y competencias necesarios para un determinado perfil académico-profesional, y que se desea que desarrollen los estudiantes que estén realizando ese tipo de estudios [1]. El programa formativo debe explicitar las competencias genéricas y específicas deseadas y distribuirlas en los cursos que configuren la titulación correspondiente, de modo de contribuir de forma eficaz y eficiente al desarrollo del perfil académico-profesional desde cada asignatura.

Aunque no hay una definición estandarizada de lo que es una competencia, puede considerarse como la integración de todos los saberes dirigida hacia una educación total del ser, basada en un aprendizaje significativo que le permita resolver los problemas que se le presenten a lo largo de la vida [2]. Las competencias representan una combinación de atributos con respecto al conocer y comprender (conocimiento teórico de un campo académico); el saber cómo actuar (la aplicación práctica y operativa a base del conocimiento); el saber cómo ser (valores como parte integrante de la forma de percibir a los otros y vivir en un contexto) [3]. Perrenoud señala que "El concepto de competencia representa una capacidad de movilizar varios recursos cognitivos para hacer frente a un tipo de situaciones" [4][5].

El ABC se fundamenta en un sistema de enseñanza-aprendizaje que progresivamente va desarrollando la autonomía de los estudiantes y su capacidad de aprender a aprender. Consiste en desarrollar las competencias genéricas necesarias y las competencias específicas, propias de cada profesión, con el propósito de capacitar al alumno sobre los conocimientos científicos y técnicos, su capacidad de aplicarlos en contextos diversos y complejos, integrándolos con sus propias actitudes y valores en un modo propio de actuar personal y profesionalmente [1]. Ello requiere el desarrollo de competencias que van más allá del mero conocimiento, y pone el énfasis en la integración entre el contenido de lo que se aprende y la estructura mental de cada estudiante, logrando que ese aprendizaje sea más duradero y significativo. Este enfoque es una respuesta adecuada de las universidades a la 'obsolescencia del conocimiento' producto de la revolución tecnológica.

En la actualidad existen diversas iniciativas que promueven la estandarización de competencias, como el caso del proyecto Tuning (*http://tuning.unideusto.org/tuningal/*), coordinado por diversas Universidades latinoamericanas y europeas, que tiene como uno de sus principales objetivos desarrollar perfiles profesionales, en términos de competencias genéricas y relativas a cada área de estudios, incluyendo destrezas, conocimientos y contenido en las cuatro áreas temáticas que incluye el proyecto (1-Competencias, 2-Enfoques de enseñanza, aprendizaje y evaluación; 3-Créditos académicos y 4- Calidad de los programas) [3][6].

La naturaleza integral del concepto competencia educativa posibilita la concreción de los cuatro pilares de la educación del siglo XXI: aprender a conocer, aprender a hacer, aprender a vivir juntos y aprender a ser [7]. Actualmente, se reconoce que las competencias propician un mayor acercamiento entre los conocimientos y el desempeño, y persiste una demanda social hacia la formación de profesionales competentes, capaces de adaptarse a los nuevos requerimientos laborales, sociales y tecnológicos, para responder positivamente a situaciones específicas y tomar decisiones que les permitan resolver problemas en forma eficaz y eficiente [8].

Ser competente es ser capaz de afrontar, a partir de las habilidades adquiridas, nuevas tareas o retos que supongan ir más allá de lo ya aprendido [9]. Demostrar competencia en algún ámbito de la vida conlleva resolver problemas de cierta complejidad, encadenando una serie de estrategias de manera coordinada.

Debido a lo anteriormente expuesto, el diseño de materiales educativos orientados al desarrollo de competencias, cobra hoy en día una especial importancia. Los Objetos de Aprendizaje (OA) son una excelente alternativa para ello, permitiendo aprendizaje abierto al desarrollo de saberes (saber, saber hacer, saber ser).

2 Modelo de enseñanza-aprendizaje centrado en el estudiante

El Proyecto Tunning latinoamericano [6] asegura que se ha llegado a un consenso respecto de las ventajas de incorporar a los procesos de formación profesional la definición de perfiles de egreso por competencias, con la identificación de los resultados de aprendizaje efectivos de los que los estudiantes deben dar cuenta a lo largo de su proceso de formación.

Este interés en el desarrollo de competencias en los programas concuerda con un enfoque de la educación centrado primordialmente en el estudiante y en su capacidad de aprender, exigiendo mayor protagonismo y compromiso por parte del alumno.

El modelo centrado en el alumno se entiende como un proceso permanente en el que el alumno va descubriendo, elaborando, reinventando y haciendo suyo el conocimiento. No propone un profesor-emisor y un alumno-receptor como el modelo tradicional, sino que el proceso aparece en una bidirección permanente en la que no hay educadores y educandos sino educadores-educandos y educandos-educadores. El profesor acompaña para estimular el análisis y la reflexión, para facilitar ambos, para aprender con y del alumno, para reconocer la realidad y volverla a construir juntos.

Como resultado de un proceso mirado de esta forma, el estudiante debe aprender a aprender. El aprender no se hace desde afuera hacia adentro, se construye internamente, y en interacción con otros, a partir de un proceso de construcción que realza el papel del propio estudiante [6]. Este enfoque implica cambios en:

- El estudiante: debe demostrar el dominio de competencias propuestas en el perfil, después de un proceso reflexivo y comprensivo de aprendizaje,
- El profesor: debe centrarse en cómo estructurar la situación de aprendizaje en función del desarrollo de las capacidades de sus estudiantes.
- La forma como se conciben las actividades educativas y la organización que se da al conocimiento: deben plantearse en función de las metas del estudiante.
- La forma de evaluar el aprendizaje: debe considerar el proceso que se ha seguido y los contextos en los que se aprende, además de los resultados obtenidos.

Es necesario tener en cuenta que las estrategias utilizadas para diseñar situaciones de aprendizaje, deben contemplar algunos principios como:

- El alumno debe participar y ser responsable de su propio proceso de formación.
- Favorecer la independencia y autonomía de trabajo.
- Permitir formas de presentación de la información adaptadas a las necesidades y características particulares del alumno.

- Hacer hincapié en los procesos de enseñanza por encima de los productos.

Además de modificar las estrategias de enseñanza, se requiere contar con materiales que posean un diseño y estructura específica, que se ajusten a los principios y metas deseadas.

3 Objetos de Aprendizaje para el Logro de Competencias

Una de las mayores preocupaciones del docente es desarrollar en sus alumnos la capacidad de integrar los conocimientos y establecer conexiones entre lo que sabe, lo que ha vivido, lo que entiende y el nuevo contenido de aprendizaje. El diseño de un óptimo proyecto de aprendizaje es un aspecto crítico al momento de garantizar la calidad de todo el proceso educativo. Se trata de estructurar y secuenciar los conocimientos propios de la asignatura y además, presentar una adecuada planificación que facilite y oriente al alumno en su proceso de aprendizaje de acuerdo a sus necesidades y disponibilidad.

Se debe potenciar el estudio independiente y autónomo del alumno y procurar que el estudiante logre las competencias con una actitud participativa y activa, obteniendo el conocimiento básico sobre los conceptos, teorías, procedimientos y técnicas propias de la materia, y desarrollando la reflexión para que puedan elaborar, interpretar y construir otros conocimientos.

Las competencias otorgan un valor agregado al proceso de enseñanza posibilitando una dinámica entre los conocimientos, las habilidades básicas y el comportamiento efectivo. Facilitar el desarrollo de competencias en el proceso de formación supone revisar las estrategias de enseñanza y de aprendizaje, de manera de garantizar que los estudiantes puedan realizar actividades que les permitan avanzar en su desarrollo. La evolución de las teorías didácticas avanza hacia un modelo educativo que se basa en quien aprende.

Los Objetos de Aprendizaje cumplen con ciertas características deseables, tales como la integración de teoría y práctica, la autoevaluación, la interacción del alumno con el material y la posibilidad de que el alumno elija el camino que seguirá para aprender, considerándose entonces un modelo de diseño de materiales adecuados a las necesidades planteadas. El proceso de enseñanza aprendizaje mediante OA permite aprender construyendo el conocimiento mediante la reflexión, la experimentación, la interacción, la solución de problemas, etc. Se considera que son totalmente adecuados para lograr el desarrollo de competencias.

3.1 Diseño de la unidad temática

En el proceso de diseño de un proyecto de enseñanza basada en competencias y mediante OA [10,11] se deben tener en cuenta las siguientes etapas para cada unidad temática:

1. Determinación de competencias y objetivos.
   Se seleccionan las competencias a alcanzar, se identifican los elementos de competencia, incluyendo la definición de las capacidades a desarrollar, así como los criterios de evaluación y se definen los saberes involucrados (saber,

saber hacer y saber ser). Se definen con claridad los objetivos a alcanzar para cada unidad temática.

2. Selección de temas
   Se establecen las relaciones y conexiones entre unidades y se consideran los conocimientos previos que pueden ser requeridos. También se identifican las situaciones problema que movilizan e integran los recursos de la competencia, los aprendizajes esperados, contenidos, actividades y evaluación del módulo formativo.
3. Armado de la Red Conceptual de la unidad.
   Se construye la red conceptual de la unidad teniendo en cuenta los conceptos involucrados y sus conexiones.

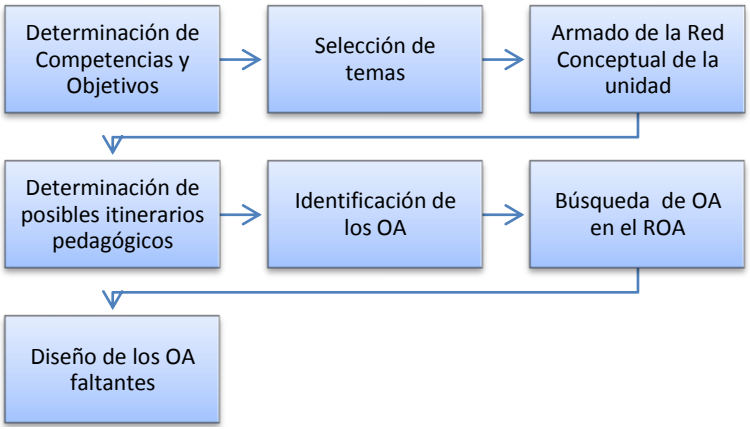


Figura 1 – Etapas del Diseño de una Unidad Didáctica

4. Determinación de posibles itinerarios pedagógicos.
   En base a la red conceptual de la unidad, se fijan las secuencias alternativas entre temas considerando las motivaciones y cortes de aprendizaje que la unidad amerita, incluyendo la evaluación, y el planteo de problemas y de casos reales [11,12]. Para contemplar los distintos cortes del conocimiento así como los intereses y los diferentes estilos de aprendizaje de los alumnos pueden construirse diferentes itinerarios pedagógicos, dando la posibilidad de múltiples trayectorias para lograr el objetivo propuesto. Algunos alumnos explorarán más caminos que otros, dependiendo de sus motivaciones y conocimientos previos.
   Algunos itinerarios posibles [12]:
   - *De conceptualización y aprendizaje*
     - *Nivel básico:* lo mínimo pretendido.
     - *Nivel medio/avanzado:* Agrega nodos o profundiza conceptos.
   - *De repaso:* Toma los conceptos principales del itinerario de nivel básico del punto anterior.
   - *De refuerzo:* Para las principales debilidades detectadas se pueden proponer reforzar algunos conceptos.

- *De ejercitación:* Para cada nodo de la red se proponen diferentes actividades que motiven la ejercitación y contemplen los diferentes estilos de aprendizaje.
- *De autoevaluación:* Cada nodo de la red puede contemplar alguna actividad de autoevaluación, como cuestionarios, selección múltiple.

5. Identificación de los OA
   Se identifican los OA más adecuados a los objetivos pedagógicos para cada contenido y se establecen las competencias a cubrir por el OA, determinando los saberes a ser tratados en cada OA. Se debe tener en cuenta que, por un lado, para alcanzar una competencia pueden requerirse varios OA, y por otro, un mismo OA puede contribuir a alcanzar más de una competencia. Desde el punto de vista de los saberes a movilizar, y en función de la complejidad, se puede considerar el desarrollo de un OA para cada criterio de evaluación de la competencia. Cada OA se basará en las capacidades expresadas en un elemento de competencia (específica o no) y en uno o más criterios de evaluación derivados de la misma.
6. Búsqueda de los OA en el ROA.
   Una vez identificados los OA, se debe ver si ya existen OA que cumplan las condiciones requeridas para ese caso y que se puedan reutilizar, para lo cual se realizan búsquedas en diferentes repositorios de objetos de aprendizaje (ROA). Si en la búsqueda se encontraran OA de interés, se debe analizar la forma en que se insertan esos OA en el modelo de enseñanza del curso. Si para utilizarlos se necesitara realizar algún cambio, habrá que verificar si es legal reusarlos o adaptarlos. Se debe revisar los permisos contemplados en el licenciamiento otorgado por el autor, si está permitida la derivación de obras nuevas a partir de la existente cumpliendo ciertos requisitos y en tal caso, si realmente se cuenta con los medios para poder modificarlos y si el esfuerzo requerido para adaptarlo, es menor que el necesario para desarrollarlo [12].
7. Diseño de los OA faltantes.
   Si no se encuentran OA ya desarrollados, se debe proceder a crearlos. Se los crea, se prueban, se evalúan, se catalogan, se empaquetan y se almacenan en el repositorio para poder ser localizados y compartidos.

Es conveniente disponer de diferentes OA relacionados con un mismo concepto, OA teóricos, OA experimentales, OA evaluativos, OA colaborativos, de manera de permitir mayor flexibilidad al momento de combinarlos y ensamblarlos para cumplir un objetivo de enseñanza, teniendo en cuenta la diversidad de los alumnos y los diferentes estilos y preferencias de aprendizaje.

Los diferentes OA que intervienen en la unidad se ensamblan para seguir cada itinerario pedagógico respetando la estrategia de aprendizaje elegida para el mismo.

El conjunto de OA que conforman una unidad de aprendizaje se puede empaquetar como una agregación de contenido SCORM para ser usado en un entorno virtual de aprendizaje (EVEA). Con el empaquetado se pueden establecer condiciones para habilitar o deshabilitar un camino en función del avance del alumno, permitiendo la personalización.

3.2 Producción de OA

El proceso de producción de los OA contempla un conjunto de acciones, resultado de un trabajo multidisciplinario de profesionales con competencias relativas a las teorías de aprendizaje, aspectos pedagógicos y metodologías de evaluación, a la temática específica, al diseño gráfico y a las tecnologías informáticas.

Si bien existen diferentes propuestas metodológicas presentadas por diversos investigadores e instituciones, la producción de OA tiene fases y etapas similares en todas ellas. Nuestra propuesta contempla:

1. Análisis y obtención del material: Consiste en establecer los objetivos y procesos necesarios para conseguir resultados de aprendizaje, de acuerdo con las competencias requeridas. Se indica claramente qué se va a enseñar, se identifican los datos generales del OA y se obtiene el material didáctico necesario para realizarlo.
2. Diseño: Se realiza en esquema general del OA, dejando en claro cómo se va a aprender. En esta fase se formulan el contenido, temario, dinámica de trabajo, sistema de evaluación, plan del curso, prácticas y actividades. Tiene el propósito de identificar y producir la forma en que se abordará el aprendizaje, se define la interrelación entre objetivo, contenidos informativos, actividades de aprendizaje y los criterios de evaluación.
3. Desarrollo: Mediante el uso de herramientas informáticas se arma la estructura del esquema general y se agrega el contenido definido en la fase de diseño. Culmina con la entrega del OA debidamente elaborado en cuanto a su estructura y funcionalidad. Es recomendable utilizar herramientas Web 2.0 que facilitan la posterior adaptación [13].
4. Evaluación: Se evalúa el OA como un todo: aspectos didácticos-curriculares (si el OA está relacionado con los objetivos, si los contenidos presentan información correcta, precisa y adecuada a los objetivos y características de los usuarios, etc.), técnicos-estética (aspectos asociados al diseño del OA: si cumple con un estándar o especificación, si la interfaz es adecuada…) y funcionales (si tipo, velocidad y nivel de interacción son adecuados, etc.).
5. Publicación en un ROA: Si el OA es evaluado satisfactoriamente, se procede a almacenarlo en el ROA elegido por la institución. Se completan adecuadamente los metadatos para permitir que sean localizados y compartidos por otros docentes.

Figura 2 – Fases de la Producción de OA

Como recurso pedagógico, un OA integra el insumo informativo, la representación para diferentes modos de percepción, el contexto de uso, el proceso o problema a resolver, las estrategias de aprendizaje, la generación de producto de aprendizaje y cualquier otro apoyo al proceso de enseñanza-aprendizaje.

Desde esta perspectiva se debe seleccionar contenido altamente significativo, vinculado con el objetivo propuesto y la o las competencias a alcanzar, crear la forma de presentación e interacción, apoyándose en las características de los usuarios o

destinatarios, tener presente posibles preconceptos de los estudiantes y construir conocimientos a partir de aquellos que ya tengan.

La información debe presentarse de forma clara, concisa y pertinente al tema tratado. Se debe ofrecer contenidos y actividades hacia las diferentes modalidades de aprendizaje (visual, auditivo), así como retroalimentación oportuna y constructiva, ejercitación que permitan aplicar los conceptos aprendidos facilitando su comprensión y aplicación a otras situaciones. Interactividad con el programa para facilitar la atención y retención de la información y potenciar el aprendizaje por descubrimiento.

Los OA de poca granularidad y alta modularidad permiten mayor flexibilidad, y por ende, mayor reusabilidad: el tema contenido en el mismo se pueda usar en diferentes contextos y relacionarse con otros OA que tratan posiblemente un tema diferente o el mismo tema desde otra perspectiva.

Además, los OA de pequeña granularidad, si bien de forma aislada pueden ser usados para facilitar un aprendizaje concreto, cuando se los une en un diseño mayor, con criterios fijos y estableciendo entre ellos una relación determinada, también pueden permitir logros que cada uno por sí solo no le sería posible de alcanzar, es decir, el todo puede ser mayor que la suma de las partes.

Con el fin de lograr OA modulares e independientes, la interfaz debe ser autónoma, sin referencias a la secuencia didáctica para que pueda ser utilizado independientemente de ésta sin merma de funcionalidad.

Es útil el uso de plantillas donde se diferencien claramente las secciones competencias, objetivo, contenidos, actividades y tipos de recursos a utilizar.

A continuación se presenta un ejemplo de la planificación de un OA correspondiente a la unidad didáctica Mamposterías, utilizado en el curso de Construcciones en Edificios e Instalaciones de la carrera de Ingeniería Civil.

3.3 Ejemplo: Plantilla para el OA Revoques Exteriores

<table>
<tr><td colspan="3">OA N° 3: Revoques Exteriores</td></tr>
<tr><td colspan="3">Descripción: Presentación de los distintos elementos constitutivos de cada ítem: revoque impermeable, grueso, fino. Cómputo de materiales.
Competencias: Administrar los recursos materiales y equipos
Subcompetencias: Ser capaz de optimizar la selección y uso de los materiales y/o dispositivos tecnológicos disponibles para la implementación. Ser capaz de valorar el impacto sobre el medio ambiente y la sociedad, de las diversas alternativas de solución.
Objetivo de aprendizaje: Capacidad de identificar los componentes de las mezclas y diseñar su constitución y cómputo de materiales.
Tipo de materiales que se utilizarán en su desarrollo: Texto explicativo. Imágenes fotográficas individuales. Imágenes fotográficas seriadas. Video.
Descripción de las actividades a incluir: ejercitación referida al diseño de los tipos de mezclas para cada función y cómputo de materiales. Autoevaluación sobre los saberes abarcados.
Contenidos:</td></tr>
<tr><td>Saber</td><td>Saber Hacer</td><td>Saber ser</td></tr>
<tr><td>Tipo y características de los materiales.
Tipos de revoques</td><td>Identificar componentes.
Calcular material necesario.
Confeccionar cómputo de materiales</td><td>Manejo cuidadoso de los equipos.
Valoración y respeto por el medioambiente.</td></tr>
</table>

3.4 Catalogación de Objetos de Aprendizaje en base a Competencias

Los OA deben estar bien catalogados para que puedan ser localizados y así el docente tenga acceso a los recursos didácticos adecuados.

Morales [14] considera que para que los OA puedan ser reutilizados según el tipo de contenido, es importante considerar una forma de clasificación a través de sus metadatos, que permita saber si ese recurso está dirigido al "saber qué", "saber cómo" o "saber acerca de". Propone una clasificación de OA y la utilización de la categoría "9-Clasificación" de metadatos LOM, para catalogar determinadas competencias estandarizadas como las que promueve el proyecto *Tuning*.

La catalogación por competencias, puede ser complementada además por un conjunto de palabras clave, a través de la categoría General de LOM. Al estar catalogados por competencias, se facilitaría la búsqueda de tales recursos. [15, 16]

Además de lo propuesto por Morales en [14, 15, 16], es importante que la descripción correspondiente a la categoría "5-Educación" de los metadatos LOM, haga referencia a las competencias que desarrolla y al contexto de aplicación, así mismo, los objetivos del OA debieran ser expresados en términos de capacidades para lograr competencias específicas.

4 Conclusiones

La tendencia actual en la enseñanza debe tener en cuenta la necesidad de aprendizaje abierto, en el cual el alumno tiene autonomía y construye su propio conocimiento. La enseñanza por competencias implica que además de adquirir conocimiento teórico, el alumno sea capaz de saber hacer (aplicación práctica de sus conocimientos teóricos) y saber ser (valores) ante una determinada situación profesional, otorgándole capacidad de resolución y de aprendizaje autónomo y contínuo.

A fin de lograr integrar estas necesidades en el ámbito educativo se requiere revisar las estrategias de enseñanza y de aprendizaje y avanzar hacia un modelo educativo que se basa en quien aprende.

Los objetos de aprendizaje son una herramienta innovadora, que contribuyen tanto por su riqueza tecnológica como por su versatilidad pedagógica, a la capacitación por competencias. La interacción del alumno con el OA y la posibilidad de que elija el camino que seguirá para aprender, estimula a utilizarlos como una herramienta adecuada para la enseñanza por competencias.

Es necesario pensar el diseño de los OA dentro de un modelo que integre las competencias con los contenidos ofrecidos y distintas maneras de aprenderlos, para que de esta manera el alumno logre los resultados esperados.

5 Referencias

[1] Villa Sánchez A. (2007). Capítulo I. Aprendizaje basado en competencia, en: Aprendizaje basado en competencias: una propuesta para la evaluación de las competencias genéricas. Universidad de Deusto. España

[2] Gonczi, A. y Athanasou, J. (1996). Instrumentación de la educación basada en competencias. Perspectiva de la teoría y la práctica en Australia. Ed. Limusa.
[3] Tuning (2007) Reflexiones y Perspectivas de la Educación Superior en América Latina. Informe Final – Proyecto Tuning –. América Latina. 2004-2007. Universidad de Deusto, España.
[4] Perrenoud, Ph. (2007) *Desarrollar la práctica reflexiva en el oficio de enseñar.* Grao. Tercera Edición, Barcelona.
[5] Perrenoud, Ph (2004) *Diez nuevas competencias para enseñar.* Grao. Biblioteca para la actualización del maestro. SEP, México.
[6] CLAR (2013) Crédito Latinoaméricano de Referencia -Universidad de Deusto Bilbao
[7] Delors, J. (1999). La educación encierra un tesoro. España. Santillana.
[8] Rychen, D. y Salganik, L. (Eds.) (2001) Defining and selecting key competencies, París: Organization for Economic Cooperation and Development.
[9] Monereo, C. (coord.). (2007). Competencias básicas. Cuadernos de Pedagogía, núm. 370, pp. 10-18.
[10] Rosanigo, Z. B., Bramati, P., y Bramati, S. (2010). Objetos de Aprendizaje para la cátedra de Proyecto I. *TE&ET | Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, ISSN 1850-9959. Nro. 5, 21-28.
[11] Rosanigo Z. B., Paur A. B. y Saenz Lopez M. S. (2010). Objetos de aprendizaje: nuevas tendencias para el diseño de materiales en entornos virtuales. Editorial: Universidad Nacional de la Patagonia San Juan Bosco. ISBN 978-950-763-099-6.
[12] Paur A. y Rosanigo Z. B.(2009). Diseño de Itinerarios: Potenciando el reuso de los Objetos de Aprendizaje. Libro de actas del XV Congreso Argentino de Ciencias de la Computación, CACIC 2009, Jujuy. ISBN 978-897-24068-4-1. Páginas 1237-1246.
[13] Chiappe (2009). Objetos de aprendizaje 2.0: una vía alternativa para la re-producción colaborativa de contenido educativo abierto. Colección: Univirtual Objetos de Aprendizaje Prácticas y perspectivas educativas ISBN: 958-8162-65-3 Pontificia Universidad Javeriana – Cali.
[14] Morales Morgado, E.M., et al. (2013). Desarrollo de competencias a través de Objetos de Aprendizaje. *RED Revista de Educación a Distancia. Nro 36. Monográfico Especial SIIE 2012. 28 de febrero de 2013.* Consultado el (10/07/2013) en http://www.um.es/ead/red/36/.
[15] Morales Morgado, E. M., Díaz San Millán, E., García-Peñalvo, F. J., (2011). Gestión de objetos de aprendizaje a través de la red, basado en el desarrollo de competencias. Teoría de la Educación: Educación y Cultura en la Sociedad de la Información, ISSN-e 1138-9737, Vol. 12, Nº. 1, 2011
[16] Morales Morgado, E. M., García-Peñalvo, F. J., Díaz, San Millán, E., Seoane Pardo, A. M. (2011). Learning Objects Searching based on Skills Development. International Journal of Computers Applications Proceedings on Design and Evaluation of Digital Content for Education (DEDCE) (2):13–19. USA: Foundation of Computer Science. ISBN: 978-93-80746-65-9

Animali@: Material educativo digital para la enseñanza de la Zoología

Martorelli Sabrina [1], Dr. Sergio R Martorelli [2], Dr. Cecilia Sanz [1],

[1]Instituto de Investigación en Informática LIDI. Facultad de Informática – UNLP,
[2]CEPAVE (CONICET-UNLP), La Plata, Buenos Aires, Argentina.

smartorelli@lidi.info.unlp.edu.ar, sergio@cepave.edu.ar, csanz@lidi.info.unlp.edu.ar

Abstract. En este artículo se presenta el diseño de un material educativo vinculado al área de la Biología y Parasitología Animal. Este constituye un nuevo avance en el marco de la línea de investigación en relación a materiales educativos digitales para estas disciplinas, y se complementa con otros trabajos ya realizados en este sentido, tales como Histologí@ y ParasitePics. El nuevo material educativo digital aborda en general las Categorías Taxonómicas, y en particular, los aspectos morfológicos y sistemáticos de los Deuterostomados, pretendiendo ser un eje más del conjunto de recursos basados en TIC, destinados a la enseñanza y aprendizaje de la Biología que se han venido construyendo. El objetivo principal para Animali@ es integrar en un material digital una serie de elementos referidos a la temática que permitirán, mediante un guión ad-hoc y consignas específicas de trabajo, complementar las prácticas de enseñanza presenciales empleadas en cursos de Biología.

Keywords: Filum Equinodermata, Material Educativo Digital Multimedia, Biología, Deuterostomados

1 Introducción

Dentro de la currícula de la materia Zoología General que se dicta en la Facultad de Ciencias Naturales y Museo, en la Universidad Nacional de la Plata, es fundamental el estudio de los seres vivos y su clasificación. El sistema de clasificación aceptado hoy en día es un sistema natural que no solo permite agrupar a los seres vivos sino que también facilita su estudio e investigación. La base de la clasificación actual la dio C. Linneo (1707-1778), que ideó un sistema jerárquico que agrupaba a los seres vivos en distintas categorías, de forma que cada categoría engloba a otras inferiores, y a su vez, se incluyen en otra superior; éstas reciben el nombre de *Categorías Taxonómicas*.

Por lo general, las clases de Biología -y de otras disciplinas- se estructuran en esquemas tradicionales. Según Acosta y Riveros (2012), se debe entender que la Biología no puede ser aprendida como un conjunto de conocimientos acabados, partiendo de haber alcanzado la verdad absoluta o como si el conocimiento fuera estático, que no evolucionara; ni mucho menos como fragmentado, tal como está sucediendo en estos momentos. Por el contrario, debe aprenderse como una integración de disciplinas (interdisciplinaridad), para que puedan combinarse puntos

de vistas diferentes. Otra condición que debe reunir la enseñanza de la Biología, acorde a estos autores, es la integración entre las clases teóricas y prácticas o de laboratorio; porque esto es una barrera para el aprendizaje, al no considerar las actividades prácticas y teóricas como procesos muy relacionados para la construcción de aprendizajes significativos (Acosta, 2012).

Con el apoyo de las Tecnologías de la Información y la Comunicación (TICs) se ha logrado romper con esa linealidad didáctica, y se busca impulsar la búsqueda de información, la construcción de esquemas y el uso de simuladores, entre otros. El uso de recursos y materiales digitales didácticos pertinentes, favorecen un recorrido propio y más integrado del conocimiento.

Bajo estas consideraciones y con la necesidad de contar con un material educativo digital en idioma español sobre dicha temática, surge la motivación para la creación de Animali@ que aquí se presenta. Este tiene como objetivo ofrecer recursos multimediales seleccionados, y otros diseñados, bajo la guía de profesionales expertos en Zoología. Al mismo tiempo, posibilitar la interactividad del alumno, explorando los diversos recursos pensados para su aprendizaje, y realizando ejercitaciones que fortalecerán su comprensión y entendimiento.

Este material ditital se complementa con *Histologi@* (Martorelli, 2013) presentado como una herramienta de apoyo para la enseñanza de la Histología Animal y con *ParasitePics*, un repositorio libre de imágenes microscópicas digitales de Parasitología Animal (Martorelli, 2012), ambos diseñado por los autores y publicados con anterioridad. Estos últimos están siendo utilizados por los alumnos de la cátedra mencionada, y en proceso de evaluación.

2 Características Generales de Animali@

Animali@ es un material digital hipermedial que se sustenta en la integración de diversos recursos dispersos en la web y otros diseñados por los autores. Su objetivo es ofrecer, a los docentes y alumnos del área, una nueva herramienta que posibilite, a través de un guión adecuado, recorrer con un hilo conductor y acorde a las necesidades de cada alumno, imágenes, videos, simulaciones y textos claves para el aprendizaje del tema en cuestión.

Animali@ se integrará por una serie de módulos que irán componiendo este material. La estrategia de diseño planificada considera ir incorporando diferentes categorías taxonómicas evolutivamente. En esta primera etapa se ha hecho foco en un grupo particular: el grupo de los Deuterostomados. Estos constituyen el mayor grupo dentro del Reino Animal que incluye a todos los vertebrados, entre los cuales se encuentra el hombre. Los Deuterostomados se encuentran divididos en tres grandes grupos: Filum Chordata, Hemichordata y Echinodermata (Tree of Life, 2013).

Si bien se pretende que el material educativo final aborde los aspectos morfológicos y sistemáticos de todos los grupos de Deuterostomados, se ha comenzado por la creación de la sección correspondiente al Filum Echinodermata. De esta manera, se podrá contar con este módulo de Animali@ para poder realizar las validaciones pertinentes, y avanzar con el desarrollo de la herramienta completa, conociendo la opinión de alumnos y docentes que aportarán para sus futuros pasos.

2.1 Descripción general del módulo Deuterostomados en Animali@

Dentro del módulo de Deuterostomados, es posible seleccionar uno de los Filums, acorde a la necesidad de los alumnos. En la Fig.1 se muestra el menú general que permite el acceso a cada uno de las secciones correspondientes a los Filums.

Fig.1. Menú General Deuterostomados

Una vez que se ha seleccionado un Filum se ha diseñado un conjunto de secciones que permitirán a los alumnos tener una ficha correspondiente a dicho Filum. La ficha se compone de los siguientes elementos, según el criterio de los docentes para la enseñanza de este tema:

- *Definición:* definición general del Filum en cuestión ubicándolo dentro del Reino Animal.
- *Características Generales:* características de la estructura anatómica en un contexto comparativo.
- *Morfología Externa:* características generales tales como la forma del cuerpo, simetría, cefalización, presencia de apéndices, etc.
- *Anatomía Interna:* forma, topografía, ubicación, disposición y relación entre sí de los órganos que componen los sistemas característicos: Digestivo, Respiratorio Nervioso, Excretor, Circulatorio, Esquelético, y Genital.
- *Aspectos Particulares de Anatomía y Morfología:* descripción de algunos aspectos particulares o novedades evolutivas que permiten caracterizar al Filum.
- *Clasificación:* división en Clases dentro del Filum con ejemplos característicos de cada una de ellas.

La ficha compuesta por estas secciones contiene una gran cantidad de imágenes, videos, simulaciones que se utilizan tanto para la ejemplificación como para la

práctica y ejercitaciones pertinentes. La ficha agrega además una sección de *Videos,* donde se agrupan recursos de este tipo, y una sección de Ejercicios que se detallará en la sección 4 de este trabajo.

3 Aspectos Técnicos de Animali@

Para el desarrollo de este material educativo digital se han utilizado los lenguajes HTML 5, CSS, una plantilla de sitios web creada con la biblioteca jQuery, el framework Mootools y la herramienta de autor Ardora.

Para la creación del menú principal que contiene las secciones correspondientes a cada Filum, se utilizó una plantilla llamada *Circle Hover Effects* (Tympanus, 2013) que utiliza transiciones CSS y rotaciones 3D para lograr un efecto de círculos animados.

Para la creación de la sección correspondiente al Filum Echinodermata, se utilizó una plantilla llamada *Moving Boxes Content* (Tympanus, 2013) la cual se encuentra implementada con jQuery (jQuery, 2013). Esta es una biblioteca de JavaScript de software libre y de código abierto. Con jQuery se logra la implementación de las animaciones que la plantilla presenta. Además, fue modificada y adaptada para el contenido específico del material.

MooTools (My object oriented tools) (Mootools, 2013) es un framework web orientado a objetos, de código abierto, compacto y modular para JavaScript. El objetivo de MooTools es aportar una manera de trabajar con JavaScript, sin importar en qué navegador se ejecute.

Puesto que es esencial para el contenido tratado, la utilización de imágenes, fotografías y videos que ejemplifiquen los temas abordados se ha utilizado Mootools para la creación de galerías multimedia a través del uso de XtLightbox (kpobococ.github.io). El uso de este complemento agiliza la implementación y mejora considerablemente la visualización de los recursos, permitiendo de una forma sencilla generar galerías de imágenes, de videos, en general, Vimeo, en particular (Vimeo, 2013).

HTML5 se ha utilizado para implementar de una manera simple la técnica de *drag and drop* para algunas actividades, que se explicarán en detalle en la siguiente sección.

Finalmente, para la implementación de cada uno de los ejercicios prácticos, se utilizó la herramienta de autor Ardora (Bouzán Matanza, 2013). Esta herramienta libre permite la generación de espacios web, a partir de diferentes plantillas para la presentación de contenido multimedia, y la creación de diversas actividades educativas, las cuales son fácilmente configurables, y pueden ser utilizadas en diversos entornos.

4 Detalles de implementación de la sección *Filum Echinodermata* en el módulo de Deuterostomados de Animalia@

La sección correspondiente al Filum Echinodermata se encuentra dividida en ocho secciones, cada una de los cuales corresponde a un ítem del menú, que aparece a la izquierda de la pantalla. Los ítems se corresponden con las secciones presentadas en el apartado de Características General de este artículo. La Figura 2 muestra la pantalla principal de la sección Filum Echinodermata.

Fig.2. Pantalla principal de Filum Echinodermata

El contenido de cada ítem se muestra a partir de un efecto de animación conseguido con jQuery. Mediante dicho efecto, se muestran pequeñas cajas esparcidas alrededor de la parte superior de la página. Cuando se hace clic en un elemento del menú, los cuadros se animan para formar el área de contenido del ítem seleccionado.

Los ítems de *Definición*, *Características Generales*, *Características Morfológicas y Clases,* presentan en su mayoría texto y agregan links a imagen y/o videos. La Figura 3 muestra cómo se presentan las *Características Generales* en pantalla.

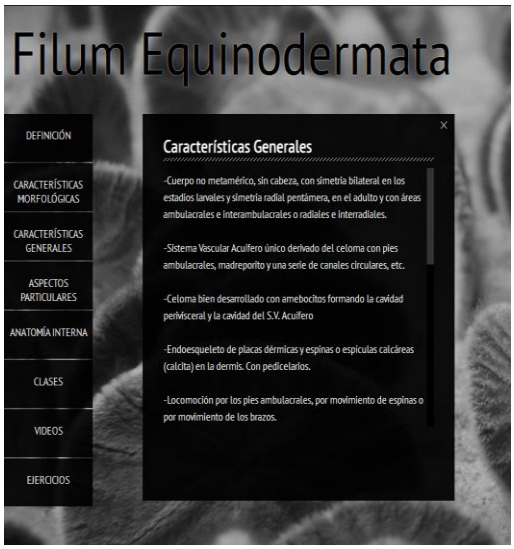


Fig.3. Ejemplo de la sección *Definición*

Para la sección *Aspectos Particulares* se ha utilizado otro efecto de animación implementado mediante JQuery. Este permite, al hacer clic sobre una imagen, que posee uno de los aspectos a ser explorado, cambiar el fondo de pantalla y presentar en él, texto correspondiente al aspecto. La Figura 4 muestra la selección de un aspecto particular y su descripción.

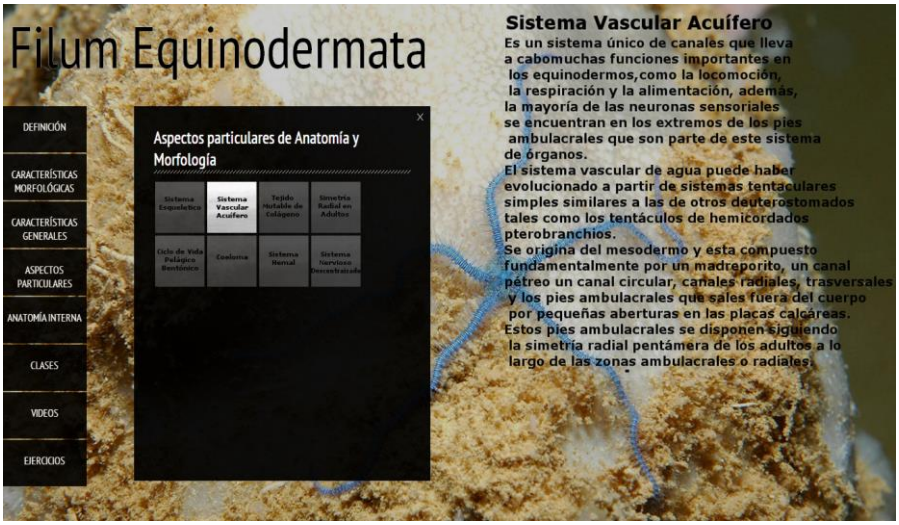


Fig.4. Ejemplo de la sección *Aspectos Particulares*

Para la implementación del ítem *Anatomía Interna,* se realizó una combinación de imágenes que mediante el uso de solapamiento de capas, y la técnica de *drag and drop* (arrastrar y soltar) permiten explorar la anatomía de un animal de una manera interactiva y novedosa. Para que esta funcionalidad pueda ser llevaba a cabo es preciso el uso de HTML 5.

Esta sección ofrece así, la posibilidad de explorar el animal que se seleccione, como por ejemplo, una estrella de mar o un erizo de mar. Al hacer clic sobre alguno de ellos se abrirá en una ventana nueva, su anatomía interna. La Figura 5 muestra la anatomía interna de un erizo de mar.

Las imágenes utilizadas para explorar la anatomía interna corresponden a una colección de imágenes de Animales Invertebrados creadas y diagramadas por Jesús Herrero Pampliega (Herrero Pampliega, 1987).

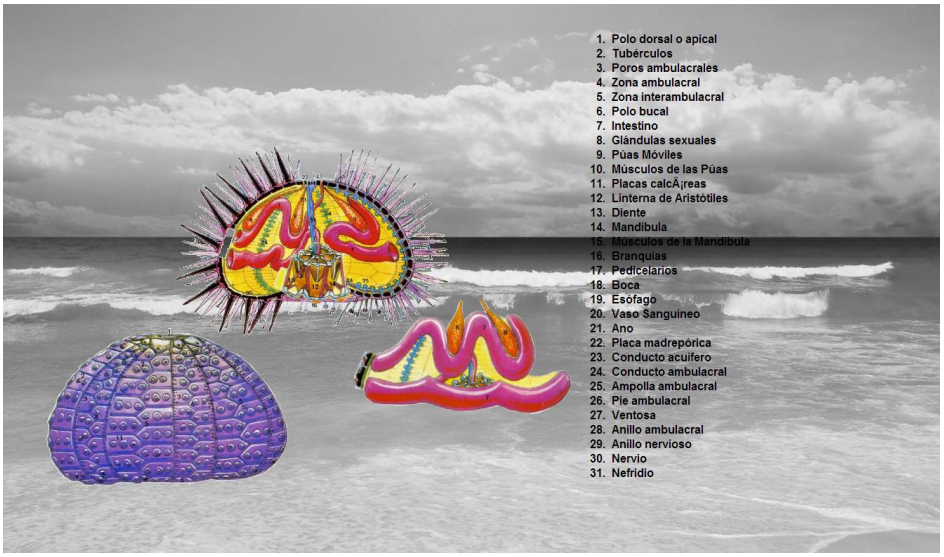


Fig.5. ítem *Anatomía Interna*

Como se explicó en la sección anterior para el ítem de *Videos*, y para mostrar todas las imágenes y fotografías asociadas al Filum, se ha utilizado XtLightbox, de manera tal de generar las galerías apropiadas. Los recursos a los que se accede se muestran en primer plano de la pantalla, en una ventada del tipo *popup,* la cual puede ser cerrada cuando se desee. Además, es posible recorrer las galerías completas con los enlaces de navegación de Siguiente y Anterior.

Antes de comenzar con la realización de los ejercicios, dentro de la sección Ejercicios, se propone la visualización de un video el cual es una adaptación de "The Shape of Life", que es una producción de la Sea Studios Foundation (Sea Studios Foundation, 2013). Dicho video aborda los temas esenciales que deben ser conocidos antes de poder realizar los ejercicios propuestos. Dicho video se encuentra en idioma ingles por lo que está siendo subtitulado por los autores.

Para cada ejercicio se decidió que se trabaje sin establecer un límite de tiempo y mostrando la cantidad de intentos o aciertos, en los casos que el tipo de actividad

permita esta configuración. Para enriquecer la devolución al momento en que se da una respuesta incorrecta, se optó por mostrar mensajes que se orienten a la motivación y a revisar nuevamente el contenido, sugiriendo la visualización del video nombrado. Un ejercicio sólo finaliza cuando se lo ha completado correctamente, y en ese momento, se recibe un mensaje de aviso de finalización.

Para la implementación de cada uno de los ejercicios se utilizó la herramienta de autor Ardora. Se trabajó sobre los siguientes tipos de actividades:

- Panel Grafico: este tipo de actividad permite al relacionar zonas en una imagen con su respectivo nombre. Se utilizó para relacionar elementos dentro de una imagen con su concepto asociado o con su nombre, de manera tal de lograr la identificación de elementos de interés dentro de cada imagen. Este tipo de ejercicio es requerido, pues se asemeja a los ejercicios que se realizan en las prácticas presenciales y permite el entrenamiento de los alumnos en la identificación.

- Seleccionar: en este tipo de actividades se ejercita el reconocimiento de elementos claves dentro de las imágenes que se presentan. Es el trabajo inverso al que se realiza en el panel gráfico, ya que en este caso la imagen no muestra las zonas a identificar, sino que deben buscarse algunos elementos que se le indican, dentro de la imagen. Al igual que el caso anterior este es un tipo de ejercicio que ayuda al entrenamiento del alumno en este tipo de reconocimiento.

- Relacionar: en este caso se utilizó una plantilla de actividad de Ardora que permite relacionar frases, o palabras, con otras fases, o palabras, en una relación del tipo varios a varios. Se trata de una actividad de relación, donde se deben asociar características generales como presencia de brazos con surcos ciliados o brazos fundidos con el disco central con cada una de las clases (Asteroidea, Ophiuroidea, Echinoidea, etc.) pudiendo estas características estar presentes en más de una clase a la vez.

- Clasificar: en las actividades de clasificación utilizadas se propone la clasificación de imágenes o palabras, según el tipo de ejercicio, acorde a la clase a la que pertenecen dentro del Filum.

- Esquemas: este tipo de actividad sirve para organizar o clasificar diversos conceptos que son presentados según un criterio específico relacionado a la temática que lleva a agruparlos. Se ha utilizado este tipo de actividad para realizar un esquema de las clases que componen al Filum.

La figura 6, presenta un ejemplo de una actividad de relación de frases. Se presenta allí una clase y debe ser relacionada con una característica propia de los organismos pertenecientes a esa clase.

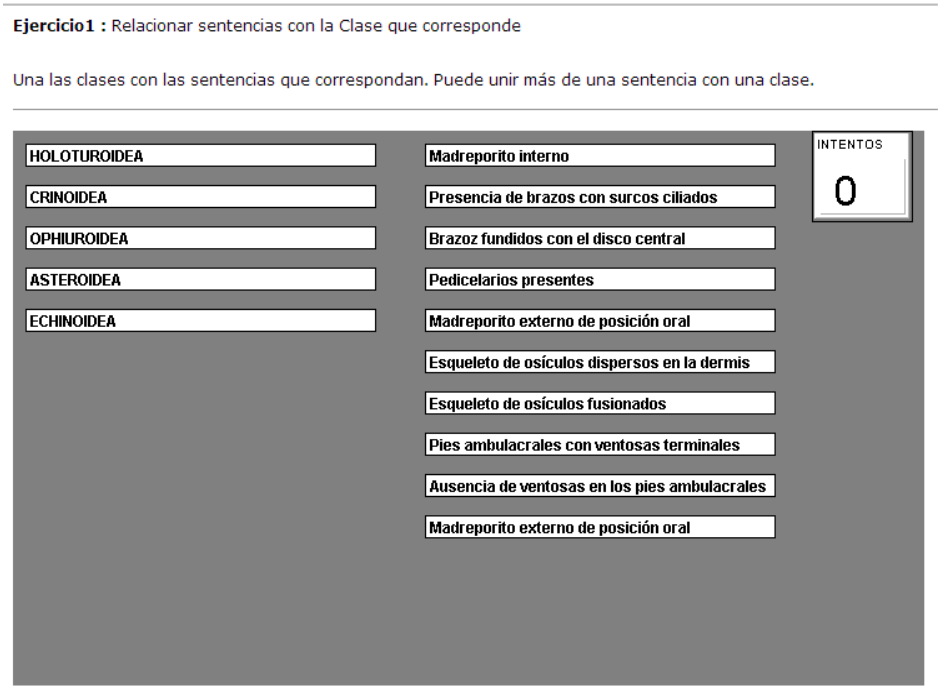


Fig.6. Ejercicio implementado con actividad de relacionar varios a varios

3 Conclusiones y Trabajos Futuros

Se ha presentado aquí el diseño y parte de la implementación de un material educativo digital, que propone presentar de una manera innovadora e interactiva los contenidos vinculados a los temas de Categorías Taxonómicas en el área de Biología. En particular, se ha iniciado por el desarrollo del módulo vinculado a los aspectos morfológicos y sistemáticos de los Deuterostomados, Filum Echinodermata.

Se espera que el material implementado sea utilizado por grupos numerosos de alumnos pertenecientes a la catedra de Zoología General de la Facultad de Ciencias Naturales y Museo de la UNLP, permitiendo una vinculación más acabada entre teórica y práctica, y un mayor entrenamiento del alumno en el reconocimiento de elementos y conceptos propios de las categorías taxonómicas.

Se considera de suma importancia en lo inmediato, y como primer trabajo futuro, realizar un primer testeo del módulo ya implementado, para que en base a la opinión de alumnos y docentes, se pueda avanzar en el desarrollo del resto del material.

Referencias

1. Acosta, Ramón y Riveros, Víctor (2012). Fundamentos teóricos para el uso de las tecnologías de la información y comunicación como mediadoras en el aprendizaje de la biología. Investigación Libre Nº 1 del Programa de Doctorado en Ciencias Humanas de la División de Estudios para Graduados. Facultad de Humanidades y Educación. LUZ.
2. Bouzán Matanza, José Manuel. Ardora 6 creación de contenidos escolares para la web. Equipo de Nuevas Tecnologías de CEIP de Palmeira. http://webardora.net. Última fecha de consulta: Julio 2013
3. Martorelli Sabrina L, Sanz Cecilia V., Javier Giacomantone, Martorelli Sergio R. (2012). ParasitePics el primer repositorio de imágenes Parasitológicas de Argentina. Revista Argentina de Parasitología (RAP) Asociación Parasitológica Argentina. ISSN 2313-9862.
4. Martorelli Sabrina L, Sanz Cecilia V., Giacomantone Javier, Martorelli Sergio R. (2012). ParasitePics: un prototipo de repositorio de imágenes de Parasitología Animal para la enseñanza y aprendizaje de esta disciplina. Proceedings del XVIII Congreso Argentino de Ciencias de la Computación (CACIC 2012) Universidad Nacional del Sur.Bahía Blanca, Buenos aires, Argentina ISBN 978-987-1648 34-4
5. Sanz Cecilia V. Ardora. (2012). Material del curso Tecnología Informática. Evolución y Aplicaciones. Maestría en Tecnología Informática aplicada en Educación. Facultad de Informática UNLP.
6. Tree of Life web project , http://tolweb.org/tree/ - Última fecha de consulta: Julio 2013
7. Maquetas recortables: Invertebrados. http://www.educa.madrid.org/portal/web/argos/invertebrados. Última fecha de consulta: Julio 2013
8. Circle-Hover--Effects-with-CSS-Transitions http://tympanus.net/codrops/2012/08/08/circle-hover-effects-with-css-transitions/. Última fecha de consulta: Julio 2013
9. Moving Boxes Content http://tympanus.net/codrops/2011/03/28/moving-boxes-content/ . Última fecha de consulta: Julio 2013
10. jQuery. http://jquery.com/ Última fecha de consulta: Julio 2013
11. MooTools - a compact javascript framework. mootools.net/ Última fecha de consulta: Julio 2013
12. Vimeo. https://vimeo.com/ Última fecha de consulta: Julio 2013
13. Herrero Pampliega Jesús. (1987). Animales Invertebrados, SENA, ISBN 9788477420057.
14. Sea Studios Foundation http://www.seastudios.com/ Última fecha de consulta a: Julio 2013
15. XtLightbox http://kpobococ.github.io/XtLightbox/ Última fecha de consulta a: Julio 2013

Diseño de una Aplicación de Aprendizaje Matemático Basada en Tecnología Android

Ruben Caceres[1] , Roy Genoff[1], Leandro Ayala[1], Patricia Zachman[1]

[1] Departamento de Ciencias Básicas y Aplicadas.- Universidad Nacional del Chaco Austral.- Argentina
eu_rubens87@hotmail.com, roy1885@hotmail.com, leansvaker@gmail.com, ppz@uncaus.edu.ar

Resumen: En los últimos años, se ha ido observando una subida exponencial del uso de los llamados smartphones (teléfonos inteligentes), así como de los hábitos de sus consumidores.
Es un hecho ya, que desde que aparecieron las potentes conexiones de datos, prácticamente todas las tareas que antes requerían del uso de una PC, se pueden ahora llevar a cabo en los smartphones. Actualmente, existen cuatro sistemas operativos sobre los cuales se basa el desarrollo de las principales aplicaciones móviles: Android de Google, iOS de Apple, Windows Phone de Microsoft y BlackBerryOS de RIM.
A pesar de tener menos aplicaciones, Google puede presumir de ser el sistema operativo más usado en dispositivos móviles.
Este proyecto se centra en el desarrollo de un pack de aplicaciones nativas para dispositivos móviles con sistema operativo Android como recurso de enseñanza – aprendizaje de matemática básica en Educación Superior. El objetivo de estas aplicaciones se enfocan en permitir la integración de la resolución analítica manual, de distintos problemas matemáticos, con la tecnología m-learning, desde procesos de formación, autocontrol y evaluación informal, en el contexto del ingreso universitario.

Palabras Clave: Apps, Matemática, Enseñanza-Aprendizaje, m-Learning, Android.

1 Introducción

El dominio de las tecnologías móviles por parte de las nuevas generaciones de estudiantes ha permitido identificar paradigmas didácticos basados en contextos de ubicuidad [1]. Se hace necesario, por ello, reconocer los cambios que inciden en estas didácticas actuales para facilitar el papel propiciador del docente en un escenario tecnológico, y, transitar hacia una concepción educativa contemporánea mediada por la comunicación informática.

Paralelamente, la Matemática se presenta como uno de los conocimientos imprescindibles en las sociedades con desarrollo tecnológico avanzado y sin embargo, la realidad pone de manifiesto que se trata de una de las áreas con mayores dificultades de rendimiento para gran parte del estudiantado universitario, observándose como una de las causas de los reiterados fracasos y deserciones

durante el ingreso educativo. [2]. La utilización de estrategias cognitivas y meta cognitivas matemáticas pareciera ser inconsistente con las heurísticas empleadas para analizar o resolver conflictos, razonamiento inductivo e intuitivo, y la comprobación de hipótesis

En este sentido, se considera que el cimiento –contenido- matemático debe fortalecerse a nivel inicial, no en el contexto axiomático de la matemática, sino en su esencia intuitiva pero formal, de forma tal que permita a los alumnos ingresantes experimentar de una manera grata y creativa "aprender - hacer matemática".

Las aplicaciones para dispositivos móviles o Apps, han sido programas pensadas y creadas para para proporcionar multitud de servicios a los usuarios de móviles. Las aplicaciones más famosas son las orientadas a redes sociales (Facebook, Twitter,…) o a servicios de mensajería (Whatsapp), pero también se destacan aplicaciones de banca online, aplicaciones de localización, aplicaciones orientadas a uso empresarial, entre otras.

En este proyecto se hizo hincapié en aplicaciones educativas, concluyendo en el desarrollo de un Pack de Apps móviles, como herramienta informal de apoyo didáctico, a estudiantes – profesores, en la autogestión y autoevaluación de soluciones a problemas matemáticos, empleando Android, plataforma de Google para sistemas móviles.

2 Contexto Universitario

La Universidad Nacional del Chaco Austral (UNCAus) presenta en su oferta académica 14 carreras de grado, para las cuales es necesario, al ingreso de cada una de ellas, completar un Curso de Ingreso en Matemática obligatorio pero no eliminatorio. En 2012 la UNCAus recibió aproximadamente 1000 alumnos ingresantes (según datos arrojados por el SIU UNCAus 2012). La situación cultural y educativa inicial de los estudiantes evidencia una heterogeneidad considerable. En consecuencia es necesario propiciar una base de partida común que garantice a los alumnos la igualdad de oportunidades, frente a la diversidad de preparación con la que egresan del Nivel Medio.

En el contexto de la UNCAus, una de las iniciativas es el Plan de Articulación entre el Nivel Medio – Polimodal y Superior, a través de cursos de nivelaciones presenciales y virtuales.

A través de diferentes estrategias inclusivas ha reconocido los cambios de paradigmas de comunicación que inciden sobre las didácticas mediadas por tecnologías, para transformar el esfuerzo educativo, centrado en la reproducción de textos, en el descubrimiento y la exploración de los contendidos para la autoconstrucción y autorregulación del conocimiento.

3 La trilogía ágil: Apps - Android - Matemática

Las Apps móviles son programas desarrollados para que funcionen en dispositivos móviles, y atiendan una tarea específica [3]. Una aplicación informática matemática móvil es un programa educativo destinado a resolver una o diferentes situaciones

problemáticas específicas del ambiente matemático, empleando como plataforma de base, la tecnología del celular.

El aprendizaje móvil (m-learning) es la adquisición de conocimiento por medio de alguna tecnología de cómputo móvil [4]. Por computadoras móviles se entiende smartphones, agendas personales digitales (PDAs), netbooks, tablet PCs y tal vez, dependiendo del tamaño, laptops.

A medida que el negocio de las aplicaciones móviles se va expandiendo y haciéndose rentable, se tienen que investigar las metodologías óptimas de desarrollo software para tales aplicaciones y entornos que lleven dicho desarrollo a éxito de una forma atractiva y eficiente. El desarrollador de aplicaciones móviles se enfrenta, además, con un escenario muy fragmentado, formado por multitud de plataforma incompatibles, como Symbian, Windows Mobile, Brew, iPhone SDK, Android, Linux o Java. Todo esto hace que el proceso de desarrollo para plataformas móviles sea más complejo.

La idea de una metodología ágil tiene dos motivaciones claras: un alto número de proyectos que se retrasan o fracasan; y la baja calidad del software que se desarrolla. La búsqueda de la solución pasa por una serie de factores: la mayor parte del esfuerzo es un proceso creativo y requiere de personas con talento, estos procesos son difícilmente planificables, modificar software es barato, las pruebas y revisión de código son la mejor forma de conseguir calidad y los fallos de comunicación son la principal fuente de fracaso.

Como se señala en [5], existen cinco factores principales que afectan a la agilidad de un proceso de desarrollo software: cultura de operación (operating culture, normas de comportamiento y expectativas que gobiernan la conducta de las personas, tanto en su trabajo como en las interacciones con los demás), tamaño del equipo de desarrollo, criticidad del software (tanto en el tiempo de desarrollo como en características específicas que tenga que cumplir el software o que vengan impuestos por los elementos donde vaya a ejecutarse), competencia técnica de los desarrolladores y, por último, la estabilidad de los requerimientos.

También argumentan que un método de desarrollo de software funciona mejor cuando se aplica a situaciones con características muy específicas, a esta clase de situaciones las llama "*home ground*" (bases) del método de desarrollo de software. En la Tabla I se puede observar la comparación entre las *bases* de los métodos ágiles y las de los procesos de desarrollo por planes o "planeados" (*plan-driven*).

Área	Metodología Ágil	Métodos Clásicos
Desarrolladores	Colaborativos, unidos, ágiles y entendidos.	Orientados al plan con una mezcla de habilidades.
Estudiantes – Profesores (Clientes)	Son representativos y se les entrega poder.	Mezcla de niveles de aptitud.
Confianza	Conocimiento tácito interpersonal.	Conocimiento explícito documentado.
Requerimientos	En gran parte emergentes y con rápidos cambios.	Conocibles tempranamente y bastante estables.
Arquitectura	Diseñada para los requerimientos actuales.	Diseñada para los requerimientos actuales y del futuro próximo.

Refactorización	Económica.	Costosa
Tamaño	Productos y Equipos pequeños.	Productos y Equipos más grandes.
Valor Premiun	Valor rápido.	Alta Seguridad.

Tabla 1. Bases para métodos ágiles y planeados (Tomado de [5])

En definitiva, el desarrollo ágil de software intenta evitar los tortuosos y burocráticos caminos de las metodologías tradicionales, enfocándose en las personas y los resultados. Promueve iteraciones en el desarrollo a lo largo de todo el ciclo de vida del proyecto. Desarrollando software en cortos lapsos de tiempo se minimizan los riesgos, cada una de esas unidades de tiempo se llama iteración, la cual debe durar entre una y cuatro semanas. Cada iteración del ciclo de vida incluye: planificación, análisis de requerimientos, diseño, codificación, revisión y documentación. Cada iteración no debe añadir demasiada funcionalidad, para justificar el lanzamiento del producto al mercado, sino que la meta debe ser conseguir una versión funcional sin errores. Al final de cada iteración, el equipo volverá a evaluar las prioridades del proyecto.

4 Metodologías Ágiles para el Desarrollo de Apps Móviles

Las metodologías ágiles poseen ciertas propiedades que las hacen totalmente aplicables al dominio del software en los móviles. La idoneidad de los métodos ágiles, como solución potencial a la elección de una metodología de desarrollo, se sintetiza en la Tabla 2.

Características Agiles	**Desarrollo para Plataformas Móviles**
Alta volatilidad del entorno	Alta incertidumbre, entonos dinámicos.
Equipos de desarrollo pequeños	Llevado a cabo por microempresas (Pymes).
Cliente identificable	Potencialmente, hay un número ilimitado de usuarios finales, pero los clientes son fáciles de determinar.
Entornos de desarrollo orientados a objetos	Java y C++
Software a nivel de aplicación	Mientras los sistemas móviles son complejos y altamente dependientes, las aplicaciones son muy autónomas.
Ciclos de desarrollo cortos	Períodos de desarrollo de 1 a 6 meses.
Sistemas pequeños	Las aplicaciones, aunque variables en su tamaño, no suelen superar las 10.000 líneas de código.

Tabla 2: Características ágiles y los rasgos y los rasgos observados en el desarrollo de software móvil

Android es una solución completa de software de código libre para teléfonos y dispositivos móviles. Es un paquete que engloba un sistema operativo, un "runtime" de ejecución basado en Java, un conjunto de librerías de bajo y medio nivel y un conjunto inicial de aplicaciones destinadas al usuario final (todas ellas desarrolladas en Java). Android se distribuye bajo una licencia libre permisiva (Apache) que permite la integración con soluciones de código propietario.

Las aplicaciones Android están programadas en Java, pero no corriendo sobre Java ME, sino sobre Dalvik, una máquina virtual Java desarrollada por Google y optimizada para dispositivos empotrados. La creación de una VM propia es un movimiento estratégico que permite a Google evitar conflictos con Sun por la licencia de la máquina virtual, así como asegurarse el poder innovar y modificar ésta sin tener que batallar dentro del JCP. Cada aplicación Android corre su propio proceso, con su propia instancia de la máquina virtual Dalvik. Dalkiv ha sido escrito de forma que un dispositivo puede correr múltiples máquinas virtuales de forma eficiente.

Es precisamente este contexto que el que dio la motivación para este proyecto:

- Desarrollar aplicaciones de software (Apps) para dispositivos móviles que permitan interactuar con los distintos conceptos en el campo disciplinar de la matemática universitaria, empleando desarrollo ágil
- Establecer un equipo de desarrolladores emprendedores que puedan crear rápidamente Apps para las distintas necesidades de la institución
- Explorar la creación de Apps en distintos ambientes de desarrollo, particularmente en el sistema operativo Android.

4.1 Mobile-D, una aproximación ideal para el Desarrollo Ágil de Apps

Mobile-D es un proyecto finlandés creado es 2005. Es una mezcla de técnicas agiles y tiene por objetivo principal conseguir ciclos de desarrollo muy rápidos en equipos muy pequeños. Se compone de distintas fases:

- Exploración: planificación, definición del alcance y funcionalidades del proyecto.
- Inicialización: identificación y preparación de todos los recursos necesarios
- Productización: en esta fase, se repite iterativamente la programación hasta implementar todas las funcionalidades.
- Estabilización: se hacen las últimas acciones de integración para asegurar que el proyecto funcione correctamente.
- Prueba y reparación**:** fase de testeo, hasta llegar a una versión estable del proyecto, según lo establecido en las primeras fases por el cliente. Se reparan errores si es necesario, pero no se crea nada nuevo.

4.2 El Entorno de Desarrollo

Android ofrece un plugin para Eclipse que extiende la funcionalidad de éste y facilita el desarrollo de aplicaciones para Android. Además, ofrece las herramientas que utiliza este plugin como scripts de ant para que puedan ser utilizados también desde otros entornos como Netbeans o Intellij IDEA15. Entre las funcionalidades de este plugin se encuentra:

• Emulador de Android. Permite elegir entre distintos terminales móviles y la versión del sistema operativo.
• El acceso a herramientas de desarrollo de Android como tomar capturas de pantalla, la redirección de puertos, la posibilidad de depurar con puntos de parada o ver el estado de las hebras y los procesos corriendo en el sistema.)
• Asistentes para la creación rápida de aplicaciones Android
• Editores de código para los distintos archivos de configuración (XML) que facilitan su comprensión y desarrollo
• Interfaces gráficas que permiten el desarrollo de componentes visualmente.

5 Mo-Math

Mo-Math (Matemáticas Móviles) es un proyecto piloto para ayudar al proceso enseñanza-aprendizaje, en el área matemática, empleando tecnología móvil.

Para la concretización del proyecto se llevaron a cabo las etapas de Mobile-D, en el marco de aplicaciones nativas.

5.1 Iniciación

Se analizó la influencia de los dispositivos móviles y aplicaciones matemáticas en el proceso enseñanza-aprendizaje, como un nuevo paradigma didáctico, por medio de una serie de aplicaciones ejecutables desarrolladas con la Metodología de Desarrollo Ágil en la plataforma Visual C# 2008. Estas Apps fueron diseñadas para ejecutarse en sistemas operativos Linux, como punto de partida del análisis. Se tuvo en cuenta que en el contexto de la Universidad Nacional del Chaco Austral, existe un gran número de docentes y alumnos que poseen computadoras personales y netbooks con sistema operativo de este tipo.

Las aplicaciones desarrolladas son sencillas de utilizar y de comprender, tanto para docentes como para alumnos.

Este primer paso permitió planificar la modelización de un sistema, aún más ágil y práctico, a implementarse en las tecnologías de los celulares.

En esta primera etapa se definió el alcance y las funcionalidades a mejorar de las aplicaciones puestas en modo de prueba.

Otro aspecto a considerar lo constituyó la plataforma sobre la cual se desarrollaría el proyecto. Finalmente, teniendo en cuenta un análisis estadístico de sistemas operativos instalados en los celulares de los estudiantes de la UNCAus, se optó por Android.

5.2 Productización

El segundo paso lo constituyó la materialización del piloto sobre tecnología móvil.

Existen variados y diversos lenguajes de programación que nos permitirían concretar el traslado de Mo-Math a dispositivos móviles. Uno de los objetivos del proyecto fue desarrollar las Apps como software libre. Java resulta lenguaje de

programación apropiado para llevar a cabo dicho proceso de traslado, ya que este lenguaje permite desarrollar software libre.

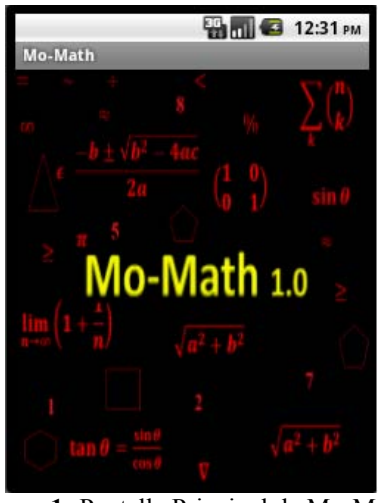


Figura 1: Pantalla Principal de Mo-Math.

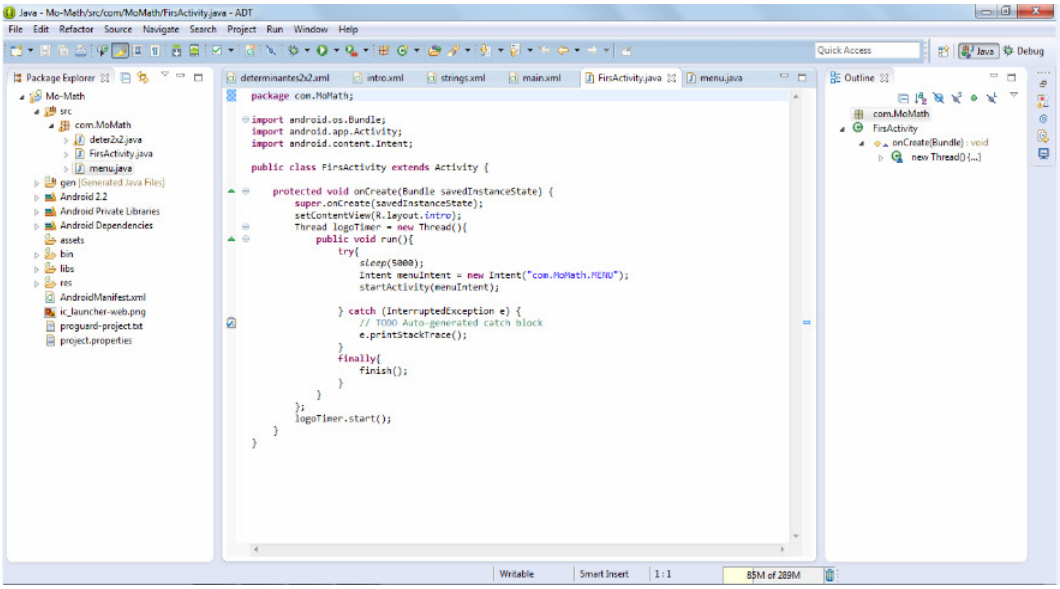


Figura 2: Entorno de Trabajo Java empleado para Mo-Math

Figura 3: Menú de Mo-Math con las diferentes Apps Desarrolladas.

En cuanto al diseño, se debió analizar si el código de adapta perfectamente o si requiere de cambios.

5.3 Estabilización

Una vez desarrollados los programas que forman parte de este proyecto, se tomó una muestra de control de 15 alumnos pertenecientes a las diferentes carreras de la UNCAus.

El proceso de enseñanza y aprendizaje de contenidos consta de partes bien diferenciadas. Primeramente, luego de las correspondientes explicaciones y demostraciones, los alumnos resolvieron analíticamente por sus propias medios diferentes ejercicios matemáticos, y seguidamente analizaron y verificaron sus resultados con las Apps móviles. Cuando el resultado obtenido por los alumnos no coincidía con los obtenidos con los programas, se planteaba un análisis sobre los errores, hasta que lograban llegar al resultado correcto. Se pudo concluir, en una primer experiencia una situación positiva: la motivación del alumno, promoviendo un aprendizaje constructivo y el autocontrol de resultados.

5.4 Prueba y Reparación

Los resultados obtenidos en la fase de prueba exigen revisar algunos aspectos como las interfaces y elementos respecto al agregado de funcionalidades.

En esta fase, el equipo se encuentra trabajando con la finalidad de refinar el producto e implementar la aplicación, en otras áreas de la UNCAus.

6 Características de Mo-Math

Existen varios software matemáticos, algunos de los cuales son complejos en su interfaz, en su forma de presentar los datos, dan lugar a confusiones debido a la gran cantidad de datos presentes en las pantallas, entradas y salidas complejas. Mo-Math presenta una serie de ventajas como por ejemplo:

1. Facilidad en la interacción con el usuario: la interacción de Mo-Math con los usuarios es fluida, no se requiere del conocimiento de comandos de uso, además con que se explique su uso una vez basta para entender su funcionamiento sencillo, solo con ingresar los datos y con un solo click se obtienen los resultados.
2. Uso de pantallas sencillas. Esta es una de las ventajas más destacadas por los alumnos de control en la etapa de prueba del sistema. Mo-Math por permitir a los alumnos una fácil interpretación de datos y resultados y también una sencilla y rápida introducción de valores que no requieren de un manual de uso ni de comandos específicos.
3. La sencilla instalación del software. La instalación de Mo-Math es sumamente intuitiva, se realiza por medio de un archivo ejecutable y lleva poco tiempo

7 Conclusión y Líneas de Investigación

El software presentado tiene la finalidad de apoyar al docente y a optimizar el proceso enseñanza-aprendizaje. Se enfatiza que estas aplicaciones están dirigidas a los alumnos del último año de la escuela media e ingresantes universitarios que deben realizar el cursillo de nivelación del área Matemática.

En cada fase del desarrollo de Mo-Math, se tuvo en cuenta el público destinatario. La sencillez de las pantallas permitió cumplir con el objetivo de ofrecer una fácil y rápida introducción de datos e interpretación de resultados. Los programas que incluye Mo-Math permiten la resolución de ejercicios y problemas matemáticos básicos del nivel medio y cubren los módulos fundamentales del área matemática: Conjuntos Numéricos, Trigonometría Plana, Expresiones Algebraicas, Relaciones y Funciones.

Este software Mo-Math, ofrece una nueva perspectiva a la forma de enseñar matemática vinculada con el uso de TICs, por lo cual, este pack de programas debe ser tenido en cuenta y sumarlo al entorno educativo de la articulación del nivel medio-universitario, para luego explotar sus potencialidades al máximo.

Se consta el hecho de que, considerando un entorno de trabajo de alta volatilidad y dinamismo, se logra establecer como elemento clave de desarrollo, el talento y la organización de pequeños equipos de desarrollo.

Desde el punto de vista práctico y más allá lo obtenido como resultados de su implementación en alumnos y docentes, resulta necesario realizar un análisis sobre desarrollo ágil para sistemas móviles que ayude a mejorar las etapas de ciclo ágil.

Los resultados obtenidos en este trabajo fueron positivos en cuanto a sus objetivos, sin embargo, los desafíos a tratar sobre temas de conectividad pueden ser más grandes y se propone trabajar algunas áreas como crear una la biblioteca de aplicaciones evaluando las próximas y actuales tecnologías o extensiones a utilizar,

generar el soporte a otros sistemas operativos como iOS (laptops, tables, smartphones y el PDA iPod touch) y sin duda alguna, generar una mayor y mejor difusión sobre esta tecnología.

8 Referencias

1. Trujillo A y Jaramillo, C Estrategias didácticas en educación superior con mediación de la computación móvil, Revista Educación y pedagogía, Enseñanza de las Ciencias y de las Matemáticas, Medllin Universidad de Antioquia, Faculta de Educación, Vol XVIII, Num 45, pp.93-107, 2006
2. Ramallo, M Panorama Actual sobre el Acceso Universitario. Revista Académica Electrònica Sementral , Vol 1 Num 1, 2012, ISSN 2314-1530
3. Traxler, J.: *Defining, Discussing, and Evaluating Mobile Learning: The moving finger writes and having writ*…International Review of Research in Open and Distance Learning, 2007
4. Traxler, J.: *Learning in a Mobile Age". International Journal of Mobile and Blended Learning,* 1-12, 2009
5. Boehm, B., Turner, R., Balancing agility and discipline: A guide for the perpexled, Addison-Wesley, 2003.

Primeros pasos en el desarrollo de ambientes virtuales inmersivos de aprendizaje utilizando software libre.

Iris Sattolo[1], Guillermo Sutz[1], Hernan Monti[1] , Jose Manuel Garcia[1] , Liliana Lipera[1]
[1]Facultad de Informática Ciencias de la Comunicación y Técnicas Especiales
Universidad de Morón, Cabildo 134, (B1708JPD) Morón, Buenos Aires, Argentina
54 11 5627 2000 int 189

iris.sattolo@gmail.com, gsutz@cnia.inta.gov.ar, manuel@latinled.com.ar, hernanmonti@gmail.com, llipera@unimoron.edu.ar

Abstract: Las nuevas tecnologías basadas en la multimedia e Internet ofrecen formas novedosas de aprender y enseñar. Una de las maneras, que hasta hace poco no existía, es la interacción mediante los sentidos de la visión, audición y tacto con los objetos y situaciones de aprendizaje, como también mediante el proceso mismo de la creación de esos objetos. Una tendencia que en los últimos años está aplicándose en las Universidades del mundo es la construcción de espacios virtuales tridimensionales en las instituciones. Por tal motivo, en esta presentación se expone el trabajo que hasta ahora se ha llevado a cabo en la Universidad de Morón, como proceso en la construcción de un metaverso que permita plantear nuevas estrategias de aprendizaje. El mismo se realizó con OpenSim como entorno libre y gratuito, permitiendo la exploración de esta herramienta para en un futuro, desarrollar espacios que faciliten la construcción de ambientes aplicando esta tecnología a la educación.

Keywords: Realidad virtual, ambientes inmersivos, Educación a distancia, Metaversos, OpenSource

1 Introducción

En este artículo se describe el trabajo de investigación que en la actualidad se está desarrollando dentro del área de inteligencia artificial aplicada al desarrollo de ambientes virtuales de aprendizaje y pertenece a un proyecto de investigación que se encuentra para su aprobación en la Secretaria de Ciencia y Tecnología de la Universidad de Morón.

1.1 Ambientes virtuales y educación.

Uno de los desafíos de todo docente, es hacer de la clase un lugar de encuentro interesante para los estudiantes, donde la motivación juega un papel decisivo en el proceso del aprendizaje, lo cual conlleva a buscar estrategias pedagógicas adecuadas.

Una estrategia que está utilizándose en los últimos años es la recreación en escenarios virtuales, en muchos de los cuales se admite la creación de contenidos propios y la interacción multiusuario mediante texto, audio y video.

A lo largo de la historia la humanidad ha experimentado innumerables cambios, nos encontramos desde las últimas décadas del siglo pasado transitando un período de avances en las tecnologías de la comunicación y la información. La sociedad de la información, como se ha llamado a esta era, ha generado grandes transformaciones y beneficios en todos los procesos, estructuras administrativas, y trabajos de las personas e instituciones involucradas. Es de esperar que en el ambiente educativo, se generen nuevas alternativas que involucren estos avances. El advenimiento de nuevas tecnologías viene acompañado del conocimiento que se genera con nueva información, pero en una sociedad de conocimiento se deben contemplar fenómenos más amplios y complejos.[1]

"La universidad y particularmente los profesores deben contribuir con una práctica educativa innovadora, para acompañar el cambio de una sociedad de información a una sociedad de conocimiento. Estos deberán adquirir ciertas habilidades y actitudes que los capaciten para aplicar estrategias innovadoras y modelos alternativos con las TIC de por medio, donde el alumno tenga un rol activo y mayor responsabilidad en el proceso de formación." (Mariño 2008).[2]

En el marco de estas nuevas tics se encuentran los ambientes inmersivos. La UNESCO (1998) [4] en su informe mundial de la educación, señala que los entornos de aprendizaje inmersivos, son una forma totalmente nueva de tecnología educativa ofreciendo una serie de oportunidades y tareas a las instituciones de enseñanza de todo el mundo. Estos Ambientes Virtuales de Aprendizaje Inmersivos (AVAI) no se dan de forma automática, ni se generan sólo como resultado de las nuevas tecnologías, el diseño pedagógico es decisivo para que surjan comunidades virtuales. Para diseñar estos ambientes de aprendizaje, se debe tener en cuenta el modificar actitudes, ideas y mecanismos tradicionales entre docentes y estudiantes.

Se puede definir a los ambientes inmersivos como entornos que permiten la recreación de escenarios tridimencionales reales o imaginarios generados por computadora con los que el usuario puede interactuar y que le produce la sensación de estar dentro.[6] Estos entornos que han sido muy utilizados en aplicaciones de entretenimiento, películas y video juegos, en los últimos años están siendo utilizados en la educación. El ambiente de educación superior es el que ha tomado la iniciativa.

Una definición de ambientes virtuales de aprendizaje inmersivos o AVAI es: plataformas tecnológicas 3D para el apoyo a los procesos de formación virtual y presencial a las cuales se accede a través de Internet o red local, permitiendo a los estudiantes y tutores conectarse para ser representados por un personaje virtual en 3D.

En este ambiente, los estudiantes pueden desplazarse libremente por los espacios de aprendizaje y comunicarse en tiempo real usando sistemas de voz y texto para realizar actividades de formación colaborativas permitiendo un nivel de interacción muy alto con los objetos de aprendizaje del entorno.

Mundos virtuales o metaversos, como Second Life, Kaneva, There, Moove, Cybertown y Active Worlds están siendo implementados desde el año 2001 y en el ámbito universitario están tomando fuerza en distintos lugares (Norteamérica**,** Europa y Asia).Actualmente, el entorno de simulación más conocido es probablemente, Second Life el cual reúne el mayor número de centros educativos y universitarios, superando a mayo de 2008 los ciento cuarenta centros, entre las que se encuentran la mayoría de las universidades pioneras, (Silva, 2009, pp. 20-21).[3]

En el año 2007 nace el proyecto OpenSim, con la propuesta de crear un servidor de aplicaciones 3D, analizando la estructura del cliente de Second Life (ingeniería inversa). Características como ser Software Libre (Licencia BSD), tener una Estructura Modular, soportar múltiples visores o clientes, y estar escrito en C# lo hacen atractivo para su uso. Esto representó para las universidades, poder construir sus propios espacios (islas) sin tener que pagar por los terrenos y tampoco por las texturas y objetos que ofrecía Second Life. [5]

En este contexto, es que surge esta pregunta: ¿Es posible que esta herramienta tecnológica, la cual crea nuevos escenarios de enseñanza-aprendizaje, acerque al estudiante-docente-conocimiento de una manera lúdica, novedosa y exitosa?

La recreación de un ambiente inmersivo en la Universidad de Morón, aplicando Software Libre, abre las puertas a nuevas propuestas dentro de la educación, en las cuales deberán plantearse nuevas estrategias de enseñanza acorde a las nuevas corrientes de pensamiento de la didáctica.

2 Desarrollo

El grupo de investigación está conformado en este momento, por Ingenieros y Licenciados en su rol de docente, y por tesistas de grado en la carrera de Licenciatura de Sistemas, aspirando a la construcción de un grupo interdisciplinario para el aporte de los requerimientos didácticos y requerimientos específicos de las materias que se propongan para comenzar las pruebas de campo.

La primera etapa planteada fue tomar conocimiento de la tecnología existente, evaluando las virtudes e inconvenientes de las mismas. Se decidió por la adopción de una tecnología OpenSource, la cual nos exime de posibles cambios de políticas en ambientes pagos. Ante las alternativas que se encontraron se decidió por el servidor OpenSimulator.

2.1 Open Sim

OpenSim es un servidor 3D de código abierto que permite crear ambientes virtuales los que pueden ser accedidos a través de una gran variedad de visores (clientes) o protocolos (software y web). *OpenSim* es un framework fácilmente configurable para cada necesidad, el que puede ser extendido usando módulos. La licencia de *OpenSim* es BSD permitiéndole ser de código libre y al mismo tiempo ser usado en proyectos

comerciales. Al día de hoy (1/7/2013), está disponible la versión 0.7.5 Existen hasta el momento dos maneras de configurar el servidor, en modo independiente o en RED. El modo independiente usa por defecto la base de datos SQLlite (base de datos ligera que no aplica persistencia), pero soporta configuraciones con las bases de datos MySql y MSSQL. [5]

La configuración del OpenSimulator consta de regiones y servicios de datos, en el modo independiente, las regiones y los servicios de datos se ejecutan en un mismo proceso. En el modo red los servicios de datos no son parte del proceso del servidor de la región. En su lugar se ejecuta un servicio llamado Robust.exe. Esto permite que se puedan ejecutar en distintos espacios físicos varios OpenSim.

La ejecución en modo red requiere tener una mayor compresión de posiciones (x,y de las regiones), contraseñas, parcelas, propietarios de inmuebles por lo que se decidió en esta primera etapa de estudio trabajar con el modo Standalone.

Desde el lado del usuario existen múltiples visores (Phoenix, Imprudence, Hippo Viewer, Firestorm, entre otros), estos son aplicaciones clientes que se instalan en sus ordenadores los cuales son el medio de poder entrar y disfrutar de estos metaversos.

3 Solución propuesta

Para poner en marcha nuestro metaverso, se optó por utilizar una máquina Pentium IV HT con 2Gb de RAM, con el sistema operativo Windows XP SP2. OpenSim trabaja con las todas las versiones de windows superiores a XP. También es posible ejecutarlo en Sistemas operativos Linux.

En una primera prueba se usó la versión Standalone con conexión a la base de datos predefinida SQLite, pero al probar la conexión con MySql se comprobó el mejor rendimiento del Simulador.

La conexión a Internet donde se encuentra alojado el servidor es de tipo hogareña con 3 Mbps de bajada y 512 Kbps. Se eligió para la instalación del metaverso la distribución OpenSim Diva Standalone. La misma tiene como ventaja una interfaz Web que permite la generación de usuarios (avatares) de forma autónoma, gestión de los mismos, gestión de regiones, obtener información del inventario. La versión instalada es Diva-r22458. [7] [8]

Se presenta en la *figura 1* el diagrama de componentes que se exploró hasta el momento: Desde el lado del cliente se accede a la página web que interactúa con el simulador y permite crear usuarios. Con los visores 3D se accede a las regiones que ofrece el mundo virtual creado.

La versión Diva ya está configurada con 4 regiones, las mismas despliegan un archivo OAR, el cual ofrece una región de inicio donde el avatar puede elegir su apariencia. En este sector está permitida la construcción de objetos por los distintos usuarios. Existe una sala de entrada con distintas aulas. Para poder construir dentro del edificio se deben obtener permisos, tarea ésta hecha por el administrador y desde consola.

Tanto la incorporación de archivos OAR o IAR se realiza desde la consola de administración. Otra tarea interesante desde consola es la de observar todos los

eventos y errores que pueden estar ocurriendo en el servidor 3D. Las configuraciones del Server, se realizan a través de los archivos de extensión INI que componen el O*pensim*.

A través de la consola además de crear usuarios, se puede modificar el terreno, enviar mensajes a todos los usuarios, establecer seguridad y todo lo referido a la administración del simulador.

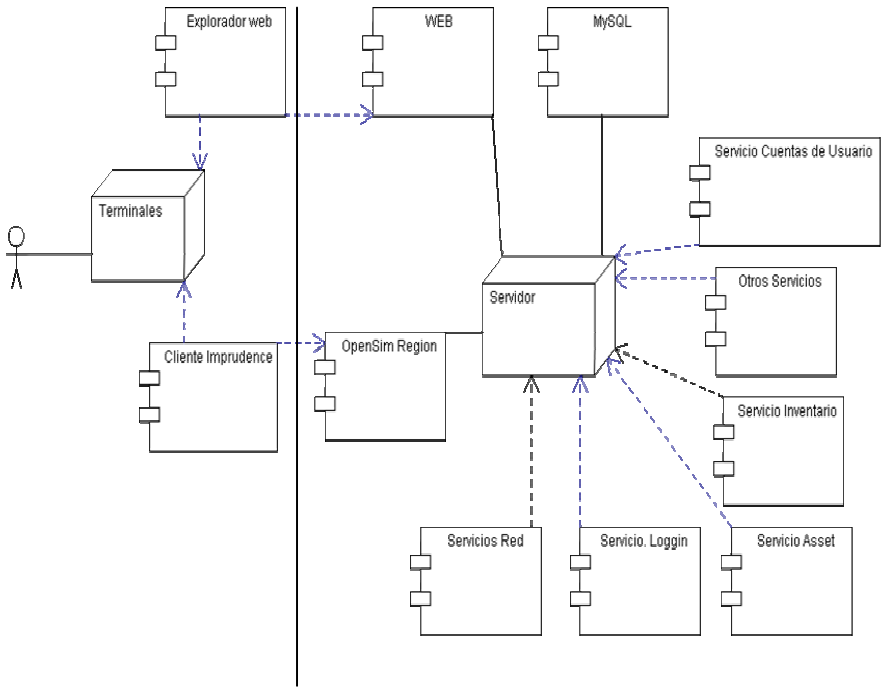


Fig.1 Diagrama de componentes de distribución DIVA modo Standalone

La distribución DIVA[7] viene con servicios de chat, pero no con servicios de voz. Se investigó sobre este punto, encontrando que la empresa que provee servicios de voz para Second Life es Vivox. La misma ofrece servicios de comunicación administrados en forma de chat de voz, mensajería instantánea (IM) y presencia en juegos, online, mundos virtuales y otras comunidades en línea. Esta empresa provee una versión gratuita para plataformas OpenSim, y está disponible para individuos, organizaciones sin fines de lucro, educadores y pequeñas redes sociales. Este servicio se solicito mediante la página que ofrece la empresa, y una vez aceptada la aprobación del proyecto, se conectó el Vivox Free Virtual World Voice Service, con el servidor OpenSim.

Las primeras pruebas de inmersión, se realizaron con avatares, a los cuales se cargaron texturas desde los visores, sin tener inconvenientes.

En una segunda fase se realizaron pruebas con los avatares configurados desde las regiones que provee el simulador. Como ya se mencionó la región de inicio provee

avatares pre-configurados con texturas provistas por la región. Esta prueba dificultaba la carga de algunos avatares, los que se visualizaban de una manera difusa con el aspecto de una nube. La *figura 2* muestra una impresión de la pantalla de esta prueba.

Fig 2. Pantalla en la cual se muestra la dificultad de carga de los avatares con texturas provistas por la región.

Se decidió que los avatares no carguen las texturas dispuestas en las regiones, sino que se vistan con la funcionalidad prevista por el visor. De esta manera no hubo inconvenientes en la visualización de los distintos avatares. Se realizaron pruebas donde hubo 6 avatares en línea, los cuales pudieron utilizar el servicio de voz previa configuración del panel en el visor. *Figura 3*.

Fig.3 Participación de 6 avatares con texturas provistas desde los visores

Se está investigando sobre el funcionamiento de los bots "programa autónomo en una red, especialmente Internet, que puede interactuar con los sistemas informáticos o

los usuarios". Para la implementación de mismos existen en la actualidad dos posibilidades, las cuales se diferencian primordialmente por programarse y ejecutarse una en el cliente y la otra en el servidor, sendas soluciones poseen ventajas y desventajas con respecto de la otra, las cuales se están evaluando para definir el tipo de implementaciones que se adoptará.

4 Resultados

En esta primera fase, se configuró un Servidor OpenSimulator al cual se accedió e interactuó entre el grupo de trabajo validando los siguientes puntos:

Libertad de acceso y de movimiento de los avatares que representan a los usuarios. Los mismos pueden moverse a través de los escenarios, caminando, dando vuelta en rededor de los objetos, volando, y teletransportándose de una región a otra.

Se detectó la posibilidad de manejar colisiones, ya que el avatar no puede traspasar paredes u objetos que denoten una estructura física.

Los objetos pueden ser animados desde la aplicación por medio de script o desde fuera de la misma.

Perspectivas de incorporar algoritmos de inteligencia artificial en la simulación de comportamientos.

Sonido espacial. El simulador permite tener en distintos ambientes sonidos propios para el lugar, comprobando que a medida que un avatar se aleja, la voz es más tenue hasta desaparecer.

Interacción de los usuarios desde lugares remotos y en tiempo real posibilitando la construcción del conocimiento del grupo sobre cualquier tema que surgiera en el momento. Esta interacción implica tener control del sistema creado.

Existió una primera aproximación al lenguaje lsl, lenguaje utilizado para desarrollar script sobre objetos, con los cuales se pueden realizar las interacciones avatar-objeto y objeto-objeto desde cualquier cliente y utilizando cualquier visor compatible. Este lenguaje esta basado en eventos, estados y funciones con una sintaxis similar al C, lo que posibilita una rápida familiarización y comprensión del mismo.

5 Conclusiones y futuras líneas de investigación

El uso de los mundos virtuales, particularmente en la educción, está siendo aceptado como herramienta a nivel mundial, acercando el conocimiento de una manera diferente a jóvenes y adolescentes, posibilitando experiencias propias sobre objetos y sobre un mundo, que no le es indiferente a los usuarios, ya que estos mundos son utilizados en mini-juegos. La posibilidad de reunir a un grupo de estudiantes, sin tener que movilizarse a un espacio físico, la posibilidad de incorporar contenidos de aprendizaje en distintos formatos, gravar esos contenidos y expandirlos a otras plataformas web, y la persistencia de los contenidos, hacen de estos mundos una herramienta que pueda ser explotada en educación. Aprendizaje, creación y exploración de modelos tridimensionales para captar la atención y sumergir al

individuo en una propuesta educativa distinta e innovadora será el desafío al que nos enfrentaremos en los comienzos de este milenio.

La expectativa de contar con un metaverso en la Universidad de Morón, en el cual se desarrollen contenidos para las distintas áreas, permitirá dar acceso a alumnos de distintas cátedras, con la posibilidad de obtener los beneficios de la educación a distancia y los de la educación presencial. Poder asistir a una clase virtual sin concurrir físicamente a la facultad, será un logro que impacte sobre los alumnos. Crear distintos espacios que representen una realidad de difícil acceso o imposibles de implementar en un entorno de educación convencional, es una meta propuesta que impacta sobre el posicionamiento de nuestra Universidad en el mundo actual.

Como primer paso en las líneas de investigación se propone crear un metaverso de modo grid, instalado en el laboratorio de la Universidad de Morón. La segunda línea en concepción trata de la creación de nuevas herramientas para aumentar la cohesión que existe entre el mundo virtual y el real, para ser utilizadas en plataformas de educación a distancia. Este mayor enlace no solo aumentará la sensación de inmersión al desdibujar la frontera entre los dos mundos sino que además serán incentivadoras para atraer a los usuarios a la utilización de esta tecnología. Entre las herramientas en consideración y desarrollo se encuentran la creación de escritorios remotos que permitan dictar clases en forma simultánea en un aula virtual con la real y el envío de mensajes de texto (SMS) desde el metaverso hacia celulares del mundo real.

6 Referencias

1. Nuevos contenidos en educación a partir del EEES de Ma Teresa Piñeiro Otero (2011)
2. Mariño, Julio Cesar Gonzalez.TIC y la transformación de la práctica educativa en el contexto de las sociedades del conocimiento rusc vol. 5 n.º 2 (2008) | issn 1698-580x
3. Silva, M. (2009). La universidad en los mundos virtuales. Educación y Mundos Virtuales Edición 24, pp. 20-21.
4. http://www.slideshare.net/RamnMartnez1/declaracin-unesco-1998
5. http://opensimulator.org/wiki/Wifi
6. http://www.revista.unam.mx/vol.8/num6/art47/jun_art47.pdf
7. http://www.marlonj.com/blog/2012/04/instalando-diva-distro-opensim-0-7-3-en-ubuntu-server-11-10/#sthash.r79wI8CD
8. http://metaverseink.com/Downloads.html consultada 5/6/2013

O Letramento E O Ensino De Literatura Mediados Por Jogos Digitais Educacionais

André Noronha Furtado de Mendonça[1],Denise Mallmann Vallerius[2],

[1]IFSUL-Pelotas, Pelotas, Praça Vinte de Setembro. 455, Centro,
96015-360Rio Grande do Sul, Brasil,
[1] PGIE/UFRGS, Av. Paulo Gama, 110, prédio 12105, sala 332, Centro,
90040-060 Rio Grande do Sul, Brasil, andrefurtadodesigner@gmail.com
[2] IFRS-Restinga, PortoAlegre, Estrada João Antonio da Silveira. 351, Restinga,
91790-400 Rio Grande do Sul, Brasil, denise.vallerius@restinga.ifrs.edu.br

Resumo.A partir da dificuldade nacompreensão de textos literáriosque alunos do ensino médio apresentam, geralmente associada a pouca valoração da leitura, desenvolveu-se uma pesquisa que explora o jogo digital como motivador do interesse pela leitura de obras pertencentes ao sistema literário brasileiro, bem como no desenvolvimento do letramento deste aluno. A pesquisa apresenta resultados ainda preliminares e dá-se por meio de revisão bibliográfica, análises quanti/qualitativas, a formação de grupos de estudo com alunos de nível médio em instituições de ensino técnico e tecnológico e a criação de um fórum de discussão disponibilizado em uma importante rede social. O emprego do jogo a ser desenvolvido apoia-se nos conceitos de aprendizagem significativa, em estudos sobre a formação social da mente, bem como no conceito de zona de desenvolvimento proximal. Os sistemas de regras, a jogabilidade e a dimensão estética do jogo serão os diferenciais abordados como elementos motivacionais do aluno/leitor.

Palavras-chave: literatura; jogos digitais; objetos de aprendizagem; ensino; estética.

1 Introdução

A partir de dados do [1], constata-se que 51,4% dos alunos matriculados em escola pública não obteve os conhecimentos esperados na avaliação de leitura da Prova ABC. O mesmo relatório aponta que, concluída a Educação Básica, menos de 30% dos estudantes no Brasil domina o conteúdo esperado em Língua Portuguesa. Na faixa entre 15 e 24 anos, o índice de analfabetismo funcional é de 15%, tornando-se cada vez mais perceptível, mesmo entre aqueles com maior acesso à informação. Os dados relativos ao baixo desempenho em leitura e em língua portuguesa talvez estejam relacionados com outros que podem

ser compilados a partir de fontes como o IBGE e o PNAD, quanto à queda no rendimento de disciplinas em diversas áreas do conhecimento e à evasão tanto escolar como universitária. Apesar dos avanços em leitura que o Brasil vem atingindo nos últimos anos, ainda figuramos no final da lista entre os países com melhor desempenho em leitura.

Sabe-se que a capacidade de ler e de compreender o que foi lido implica em independência e desenvolvimento, tanto do indivíduo como de toda a nossa sociedade. O indivíduo que compreende o que lê está plenamente capacitado a inserir-se em seu tempo e tem a liberdade de decidir seu agir de acordo com sua própria interpretação de mundo.

Destarte, o desenvolvimento de novas metodologias pedagógicas que instiguem práticas de leitura entre os jovens torna-se extremamente necessário, e é nesse sentido que se insere o jogo digital a ser desenvolvido pela pesquisa aqui apresentada, uma vez que se objetiva lograr um jogo educacional que sirva como ferramenta de aperfeiçoamento do letramento crítico do aluno e como motivador do interesse pela leitura de obras de nosso sistema literário.

Uma das mais importantes características do jogo (digital ou convencional) é o fato de o ato de jogar ser um ato voluntário, ainda que promovido por algum tipo de apelo social. Tal característica torna, por si, o jogo digital um potencial instrumento pedagógico.

1.1 Hipótese De Pesquisa

A aplicação do jogo a ser desenvolvido visa investigar a influência da dimensão estética dos jogos digitais como meio de estimular/engajar o interesse pela leitura em estudantes do ensino médio matriculados na rede pública de ensino. A aplicação do jogo agiria como um elemento mediador de pré-leitura, estimulando o desenvolvimento crítico e cognitivo do aluno em sua ZDP, baseando esse método, em procedimentos propostos por [2], [3] e [4].

De acordo com [2], [3], por exemplo, sugere uma ênfase ainda maior na importância do aprendizado da linguagem escrita, que para o autor, representa um salto considerável no desenvolvimento pessoal do indivíduo. Para [2], [3] a linguagem escrita é responsável por construir processos psicointelectuais inteiramente novos e complexos na mente humana, principalmente na criança. [2], [3] ressalta a característica complexa da linguagem escrita por ser constituída por símbolos de segunda ordem, onde o símbolo da linguagem escrita se articula como

representação ou designação dos símbolos verbais. Para o autor, a leitura é uma atividade, cujo domínio de seu complexo sistema de signos, é capaz de produzir o incremento na capacidade de pensamento, ao multiplicar a capacidade de memória ou registro de informações, propicia novos caminhos para a organização do agir humano e o acesso ao nosso acervo cultural. Considerando que a palavra escrita aja como um repositório do conhecimento humano capaz de vencer barreiras espaciais, temporais e históricas, exponencialmente superiores às da língua falada, parece ser naturalmente aceitável que o desenvolvimento de ferramentas que estimulem no jovem, em especial, aqueles cuja origem social seja a de classes menos privilegiadas, se torne uma contribuição importante para a nossa sociedade.

1.2 Objetivos

- Incentivar os alunos do ensino médio a terem o hábito da leitura de textos literários;
- Desenvolver um objeto de aprendizagem (jogo digital educacional), que faça uso da linguagem estética, da narrativa e da mecânica de jogo como características de projeto capazes de engajar nos alunos do ensino médio da rede pública federal no interesse pelo próprio jogo e, em consequência, pela leitura;
- Aperfeiçoar a capacidade leitora/interpretativa, dando sentido à leitura realizada por esses alunos, o que repercutirá positivamente em todas as outras áreas do conhecimento pelas quais eles transitam ou vierem a transitar.

2Letramento Critico E Jogos Digitais Educacionais

Para iniciarmos a discussão sobre a relação entre letramento e jogos digitais, é importante definirmos a diferença entre alfabetização e letramento. A alfabetização é a aquisição do conhecimento sobre o uso do código da língua escrita e o letramento é um fenômeno social e político na relação do indivíduo com a sociedade na qual está inserido, usando a língua escrita como instrumento de cidadania [4], [5] e [6].

A linguagem é um elemento complementar e fundamental na constituição da natureza humana [2], [3]. Será por meio da linguagem que o ser humano irá construir os significados simbólicos, os sistemas

de signos construídos ao longo da história e que formam os elos sociais com outros elementos de seu grupo, bem como com o mundo no qual o indivíduo está inserido. A linguagem funciona como um mediador de signos, sendo portadora e transmissora de conceitos e o meio pelo qual se possibilita o desenvolvimento dos processos psicológicos fornecidos pela cultura.

Os seres humanos criam instrumentos e sistemas de signos cujo uso permite transformar e conhecer seu próprio mundo, comunicar experiências e desenvolver novas funções psicológicas [2], [3]. Esse processo ocorre por meio do diálogo, ou, em seu sentido mais amplo, por meio das interações sociais. Para [2], [3], a aprendizagem ocorre em todos os espaços concebíveis, muito além daqueles compreendidos dentro da escola, e, nesse sentido, associando conceitos como interação social e linguagem enquanto meios catalizadores da apreensão de conhecimento, os jogos, digitais ou não, tornam-se terremos férteis, com grande potencial em termos de conformarem-se em veículos promotores de aprendizagem significativa. Essa visão de aprendizagem múltipla e continuada coincide com o trabalho apresentado por [7], que destaca a necessidade de o aprendizado ser consolidado e absorvido por diversas vias.

Existe uma continuidade entre as diversas atividades simbólicas (os gestos, as imagens e os jogos, ou brinquedos) [2], [3], contribuindo para o desenvolvimento de representações simbólicas. É a articulação dessas atividades que proporciona o processo de aquisição do aprendizado da linguagem escrita. O termo "brinquedo", aplicado por [2], [3], refere-se ao ato de jogar e compreende o ambiente no qual o aprendiz exercita, ao se envolver em situações imaginárias, suas capacidades de representação simbólica, encontrando as motivações internas para atuar em uma esfera cognitiva.

Para [2], [3], existem três níveis de desenvolvimento humano: o primeiro é o desenvolvimento real, relacionado com as capacidades ou funções já apreendidas ou dominadas pelo indivíduo. O desenvolvimento real implica em ações ou raciocínios gerenciados e produzidos de forma natural pelo indivíduo, sem que haja a necessidade de qualquer assistência por parte de algum tipo de tutor. O segundo nível de desenvolvimento é o potencial, que se relaciona com as capacidades potenciais que o indivíduo pode realizar mediante orientação e acompanhamento de um tutor, do diálogo, da imitação, da leitura e da observação visual ou discursiva sobre algum conhecimento.

O terceiro nível de desenvolvimento refere-se ao processo de consolidação do desenvolvimento potencial em desenvolvimento real, o que ficou conhecido como zona de desenvolvimento proximal (ZDP). A ZDP é, então, o processo de amadurecimento de um determinado conhecimento.

O letramento crítico, imerso como processo na linguagem, implica no desenvolvimento cognitivo do sujeito e emprega práticas sociais que usam a escrita para atingir essa meta [4]. Nesse sentido, desenvolver o letramento crítico através do emprego de jogos parece guardar uma inter-relação com o modelo previsto por [2], [3], quando enfatiza a importância do aprendizado da leitura e da escrita como veículos potencializadores do desenvolvimento cognitivo das capacidades cerebrais superiores, tais como o raciocínio abstrato, a memória ativa e a resolução de problemas, entre outros.

[4] observa que o modelo atual de ensino de língua portuguesa e literatura provoca sistematicamente o bloqueio do desenvolvimento da ZDPem alunos do ensino público e privado em nosso país. Os livros didáticos e os livros de literatura passam a ser encarados pelo aluno apenas como fontes de informação e não com fontes de uma leitura curiosa e prazerosa. Com relação à curiosidade e aoprazer, [4], apontaque é necessáriotrabalharcom o alunodoisaspectosfundamentais: a afetividade e a valoração. Para aautora, essesdoisfatoresirãodeterminar a qualidade da interação do aluno com a escrita e com seumundo. A autoraaborda o termo "afetividade" comoumarelação de confiança entre o aluno e o professor e entre o aluno e o conhecimentoquedeveráseraprendido. É precisoque o alunoperceba o valor queaescritarepresentaemsuaformaçãohumana. Talvezessaseja a contribuiçãomaisrelevante no trabalho de [4] – sem a valoração do objeto a seraprendidonãoocorre a aprendizagem. É verdadeque o PrazeridentificadoemVigotsky e que o AfetoemFreire, proporcionamcondiçõesférteisaoaprendizado, mas se o alunonãoperceber o valor do queseráaprendido, a aprendizagemnãoocorrerá, pormelhores e maisfavoráveisquesejam as condiçõesapresentadaspeloobjeto de aprendizagem.

A autora sustenta que é necessário reverter essa tendência com métodos de ensino que sejam capazes de engajar novos leitores. Por outro lado, como a própria autora observa que o aluno que frequenta a escola pública, via de regra, não possui familiaridade com a linguagem

que ele encontrará em livros didáticos ou em livros que pertençam ao nosso sistema literário, em geral, mais próximos ou pertencentes à língua culta. Nesse sentido, talvez um jogo digital que apresentasse ao aluno uma temática de jogo baseada na complexidade de uma obra de nosso sistema literário também encontrasse a mesma barreira de entendimento, não provocando no aluno o engajamento esperado. Em seu trabalho, a autora desenvolve a questão da barreira da linguagem e procura apresentar um método próprio que viabilize a aproximação desse aluno com a leitura. O método de [4]envolve uma apresentação pré-textual dos textos a serem lidos pelos alunos através de oficinas, práticas em grupo, discussões e a apresentação das histórias em momentos que antecedem a leitura do texto propriamente dito. O foco da autora não enfatiza as obras literárias, mas a compreensão de textos do dia a dia como a análise de matérias em revistas e em jornais. Pretendemos seguir o caminho proposto por [4], com a aplicação dos jogos digitais como um recurso complementar, capaz de reduzir a rejeição ao livro ao amplificar o contato prévio com o universo literário, traduzindo-o de forma imersiva e lúdica para o aluno da escola pública.

3Resultados Parciais

Este artigo apresenta resultados ainda preliminares, uma vez que o jogo digital educacional que é o objeto de trabalho em nossa pesquisa ainda se encontra em fase de desenvolvimento.

O levantamento de dados consistiu em quatro frentes principais: a pesquisa bibliográfica em artigos, trabalhos científicos, livros, revistas na área de jogos digitais e *sites* dedicados à análise, divulgação e avaliação do mercado de jogos; a análise do mercado norte-americano de jogos digitais através do anuário [10]; a organização de grupos de estudo que fornecessem dados quantitativos e qualitativos a partir da experiência com jogos digitais educacionais já existentes; e a criação de fórum de discussão que permitisse observar e conhecer melhor o perfil do jogador brasileiro.

Quando iniciamos a leitura do referencial bibliográfico, houve uma indicação consistente sobre a influência que uma estética mais sofisticada, como a cinematográfica, imprimia no sucesso de um bom jogo digital. O cruzamento de dados fornecido pelo anuário [10] entre

os anos de 2008 e 2011 também contribuiu para confirmar tal hipótese. Nesse sentido, o sucesso do jogo a ser desenvolvido dependeria também do emprego de uma dimensão estética sofisticada. Por outro lado, a revisão bibliográfica sobre a teoria dos jogos sustentava que o sucesso de um jogo depende da qualidade de seu sistema de regras e de sua jogabilidade [11]. O confronto dessas duas tendências forçou a investigação de como se comporta o público-alvo de nosso objeto de pesquisa, por meio da organização de fórum de discussão e de grupos de estudo.

Para iniciar o desenvolvimento desse projeto, foi necessário, primeiramente, identificar se havia algum jogo no mercado comercial ou educacional com uma proposta semelhante à que pretendemos desenvolver e que pudesse servir-nos de parâmetro. No levantamento de dados realizado, identificamos o recente lançamento (2013) de um projeto de jogos para dispositivos móveis desenvolvido por uma grande empresa de telefonia. A proposta desses jogos é fazer com que os usuários tenham contato com obras do sistema literário brasileiro - algo muito próximo ao que estamos propondo. Também identificamos dois jogos, premiados nos últimos anos, com uma proposta de narrativa histórica que, acreditamos, possam servir como modelo no desenvolvimento de nosso jogo. Em um segundo momento, procuramos identificar e mapear o comportamento do perfil de usuários de jogos digitais no Brasil.

Para avaliar os jogos educacionais desenvolvidos para dispositivos móveis foi organizado um grupo de estudo com 20 alunos de ensino médio de uma escola da rede de ensino federal no Rio Grande do Sul. Os jogos desenvolvidos para a empresa de telefonia foram apresentados para os participantes desse grupo que, após uma fase de imersão nos jogos, foi submetido a um questionário que visava compreender as impressões desses voluntários sobre os jogos apresentados.

Para avaliar os jogos digitais educacionais com narrativa histórica, foi organizado outro grupo de estudos na mesma instituição de ensino, desta vez, formado por 46 alunos. Os alunos passaram por uma fase de imersão nos jogos e, posteriormente, responderam a um questionário que visava identificar a eficiência desses jogos quanto ao engajamento dos alunos na busca por mais informações sobre os aspectos históricos apresentados pelos jogos, além de avaliar a qualidade do conhecimento apreendido pelos alunos.

Para que obtivéssemos informações sobre o perfil do jogador brasileiro, foi desenvolvido um fórum de discussão disponibilizado *on-line* em uma das maiores redes sociais existentes no Brasil. Neste fórum, disponibilizou-se uma enquete composta por treze perguntas, cujas respostas possuíam caráter quantitativo e qualitativo. O fórum é aberto à adesão de novos participantes e, atualmente, é composto por 550 membros, em sua maioria dos estados do Rio de Janeiro, Rio Grande do Sul e São Paulo. Seus participantes têm liberdade para escolher as perguntas as quais desejam responder, e de contribuir com comentários e/ou testemunhos pessoais, narrando suas experiências de jogo e preferências com relação a gêneros de jogo, ou qualquer outro assunto relacionado a jogos digitais.

Observando de uma forma geral as respostas fornecidas pelos alunos que participaram dos dois grupos de estudo formados, bem como os resultados obtidos por meio das perguntas do fórum sobre jogos digitais, foi possível elaborar uma síntese definindo as diretrizes de projeto que auxiliarão no desenvolvimento do jogo digital educacional que propomos:

- Ao contrário da tendência evidenciada pela evolução histórica dos jogos digitais comerciais, um jogo digital não precisa necessariamente apresentar qualidade cinematográfica para proporcionar interesse e entretenimento lúdico.
- Jogos de apontar e clicar não são necessariamente menos atraentes que jogos de ação/aventura, desde que proporcionem uma experiência de exploração do ambiente e estimulem o pensamento lógico do jogador.
- Jogos que apresentem uma história envolvente, interessante e que possuam personagens cativantes podem oferecer qualidade gráfica menos refinada sem produzir grandes prejuízos na imersão do jogador.
- Jogos que combinem narrativa envolvente, conteúdo histórico, raciocínio lógico, exploração de terreno e resolução de problemas ajudam a despertar o interesse do jogador em se aprofundar em leituras complementares para melhorar sua performance de jogo.
- Jogos com mais liberdade no movimento do personagem geram maior interesse no jogador.
- Jogos em 3D são mais atraentes que jogos em 2D, porém a diferença não se mostrou significativa.

- O realismo na representação gráfica não foi considerado importante, mas a verossimilhança é desejável. Por exemplo, se existe um obstáculo ”físico”, ele deve impedir o avanço do movimento do personagem, como ocorreria se fosse no universo real - a não ser que o personagem possuir poderes mágicos fizesse parte do jogo, o personagem não deveria poder atravessar objetos sólidos ou outros personagens.
- Quantidade excessiva de texto no jogo, pode provocar rejeição deste pelo jogador.
- A inserção de legendas informativas e mapas de localização aumentam o conforto do jogador e reduzem a rejeição ao jogo.
- Falhas de programação e a exigência de processamento de imagens que provoquem o travamento das máquinas durante a experiência de jogo aumentam o grau de rejeição por parte do jogador.
- O jogo deve contemplar um equilíbrio entre a capacidade de exploração do ambiente, a qualidade gráfica e sonora dos cenários, as capacidades de mobilidade do personagem e a capacidade de processamento do computador onde o jogo está instalado.
- O jogo deve permitir salvar uma fase para reinício futuro.

4Estágio Atual De Desenvolvimento

Atualmente o jogo digital educacional que propomos ainda se encontra em fase de desenvolvimento. O estágio atual é o de finalização do roteiro do jogo e a elaboração dos fluxogramas das fases do jogo, do roteiro e das cenas, refinamento das missões de jogo e da árvore de decisão.

Como apresentado no item anterior, a pesquisa com os grupos de estudo e com o fórum de discussão sobre jogos digitais forneceu dados para compormos um conjunto de características importantes a serem consideradas ao longo do desenvolvimento do jogo digital educacional que propomos. Por outro lado, também se verificou a necessidade de buscar fundamentação técnica e teórica que tornasse possível esse desenvolvimento. Com relação ao sistemas de regra do jogo, tipificação e qualificação de gênero, desenvolvimento da história e narrativa de jogo, mecânicas de jogo e jogabilidade, utilizamos os conceitos e diretrizes apresentados por [9], [10] e [11]. Com relação ao desenvolvimento dos personagens e ao desenvolvimento do *storyboard*

das cenas, utilizamos os modelos propostos por [10]. O roteiro do jogo foi baseado nos modelos propostos por [10] e [11]. O design do jogo e os aspectos estéticos tiveram como base os autores [11] e [12].

Por questões como limitação do tempo de produção, de equipe e de recursos financeiros, além da necessidade de levarmos em conta os equipamentos onde o jogo deverá ser instalado, definiu-se que o jogo será desenvolvido apenas para a plataforma MS Windows. Para reduzir as exigências de processamento e sistema, abdicou-se do uso de ambiente 3D real. Em vez disso, o espaço tridimensional será simulado e rodará no formato SWF. O jogo não será multiusuário, mas conterá elementos comuns a diversos gêneros, como características de ação/aventura, tiro, perspectiva em primeira e em terceira pessoa, exploração de cenários e resolução de problemas, entre outras características.

Referências

1. Anuário Brasileiro da Educação Básica, http://todospelaeducacao.org.br/biblioteca/1450/anuario-brasileiro-da-educacao-basica
2. Vigotsky, L. S. , Luria,A. L., Leontev, A. N.: Linguagem, Desenvolvimento e Aprendizagem. 12 th ed., pp. 103--117 and 191--224.Icone Editora, São Paulo (2012)
3. Vigotsky, L. S.: A formação social da mente. 1 st ed., pp. 3--20, 94--124 and 154--165. Martins Fontes, São Pulo (2007)
4. Terzi, S. B.: A Construção da Leitura. 2 nd ed. Pp. 46—144. Editora Pontes, Campinas (2006)
5. Tomasello, M., Berliner, C. Origens Culturais da Aquisição do Conhecimento Humano. Martins Fontes, São Paulo (2003)
6. Torquette, A. O. Representações Sociais de Escrita e de seu Ensino para Alunos do Esino Médio. Programa de Pós-Graduação em Estudos Linguísticos – Ensino Aprendizagem de Línguas, Universidade Estadual de Maringá, Paraná, Brasil. Dissertação de Mestrado (2009)
7. Thomas, M.: Contextulizing Digital Game-BasedLanguage Learning: TransformationalParadigm Shift or Business as Usual?. In: H. Reinders (Ed.).: Digital Games in Language Learning andTeaching, pp. 1--21. PalgraveMacmillanPublishers, London (2012)
8. 2012 Sales, DemographicandUsage Data: FactsAboutthe Computer andVideo Game Industry, pp. 4--16. ESA – Entertainment Software Association (2012)
9. Salen, K., Zimmerman, E.: Regras do Jogo, vol. 1, 2, 3 and 4. Editora Blucher, São Paulo (2012)
10. Chandler, H. M.: Manual de Produção de Jogos Digitais, 2 nd ed., pp. 139--256 and 327--338. Bookman, Porto Alegre (2012)
11. Rabin, S.: Introdução ao desenvolvimento de games, 2 nd ed., vol. 1, pp. 33--150. Cengage Learning, São Paulo (2011)
12. Marsal, M. A. A.: Teoria e Conceitos para uma Mídia Indisciplinada. Programa de Pós-graduaçãoem Ciências da Comunicação, Universidade do Vale do Rio dos Sinos, São Leopoldo, RS, Brasi. Tese de Doutorado (2011)

Una Aplicación Móvil para el Museo de Física de la Universidad Nacional de La Plata

F.Javier Diaz[1], Ivana Harari[1], Andrea Gallego[1] y Leandro Aguilar[1]

[1] Facultad de Informática, Universidad Nacional de La Plata.
50 y 120, La Plata 1900, Buenos Aires, Argentina
{jdiaz, iharari}@info.unlp.edu.ar
andreamgallego@yahoo.com.ar
leandrotaguilar@hotmail.com

Resumen. Hoy en día junto con los avances tecnológicos surgen nuevas formas de interacción, afectando la forma de visualización y de comunicación que se diseña para los usuarios. El surgimiento de los dispositivos móviles con sus sensores, cámara, capacidades gráficas y de procesamiento, da lugar a la posibilidad de desarrollar aplicaciones donde la interacción con el usuario sea optimizada y adaptada. En este sentido, se trabajó en un proyecto entre la Facultad de Informática y el Museo de Física, ambas instituciones de la Universidad Nacional de La Plata (UNLP), que consistió en el desarrollo de una aplicación móvil que utiliza código QR para mejorar la visualización de los instrumentos en exposición y proveer formas no tradicionales de recorrido y visitas del museo. Enriqueciendo de esta manera, con información aumentada, sintetizada y presentada en distintos formatos multimedia, el patrimonio cultural que allí se conserva.

Palabras claves: Tecnologías de la información y comunicación TICs, Código QR, Realidad aumentada, Sistema operativo Android, Aplicaciones móviles.

1 Introducción

El museo de Física de la Universidad Nacional de La Plata (UNLP), es un espacio de encuentro con la ciencia abierto a todo el público. En su interior, alberga una colección de más de 2.000 instrumentos utilizados para la enseñanza de la Física en las Universidades de principios del siglo XX. Ofrece una atractiva propuesta a través de sus importantes instrumentos centenarios de demostración de fenómenos físicos para todos aquellos interesados en acercarse al conocimiento de la Física.

El delicado equilibrio entre museo histórico y centro interactivo de ciencias obliga a minimizar el uso del acervo, buscando opciones que posibiliten un acceso a mayor cantidad de personas a través de la utilización de nuevas tecnologías dentro de la exposición y la difusión masiva a través de la Web [1]. Esto posibilita optimizar la observación presencial como virtual de la mayor cantidad de instrumentos en funcionamiento y sus características.

Una de las líneas que se realizaron en este marco de modernización tecnológica del museo y que se va a detallar en este artículo, es la utilización de dispositivos móviles para que, mediante el código de respuesta rápida (Código QR) [2], permita al visitante del museo amplificar la vista y observación de los instrumentos en exhibición, con el aumento de la información de los mismos. Este aumento en la percepción del instrumento por parte del usuario, se manifiesta a través de recursos multimedia adicionales, como ser detalles de imágenes ampliadas, videos del instrumento en funcionamiento, audio con sonidos y explicaciones que lo identifican, juegos educativos específicos, que se ponen de manifiesto a través del dispositivo móvil.

Modernizar las muestras itinerantes con información ampliada de la sala y el patrimonio del museo a través del uso de los dispositivos móviles, es una propuesta atrayente que no sólo permite estrechar el acercamiento con los niños y jóvenes, sujetos asiduos a la tecnología, sino también facilitaría el acceso a la información de personas con discapacidad que concurren a las visitas del museo regularmente, dando lugar a recorridos auto gestionados y a experiencias enriquecedoras.

2 El Museo de Física

El Museo de Física de la UNLP, es una institución que pertenece al Departamento de Física de la Facultad de Ciencias Exactas [3]. Está dedicado a la exposición, investigación, experimentación y divulgación de actividades educativas cuya misión es promover el acercamiento a la ciencia y la tecnología, en un ambiente atractivo e informal. Constituye un vínculo entre la extensión y la docencia formal [1].

El museo ofrece un sistema de visitas guiadas que está dirigido al público en general y a grupos de nivel preescolar, escolar, terciario, universitario. Concurren también grupos de personas pertenecientes a escuelas de educación especial.

Este establecimiento cuenta con una sala central con vitrinas sectorizadas y clasificadas según las diferentes áreas de la Física como ser óptica, mecánica, ondas y sonido, electromagnetismo y astronomía. Estas vitrinas albergan instrumentos centenarios que son mostrados y explicados de una manera didáctica y pedagógica por parte de un equipo de profesionales de diferentes disciplinas.

En un principio, los objetos dentro de las vitrinas como las vitrinas en sí, no poseían ningún cartel o información que los identifique y que explique su funcionamiento, dificultando al visitante la posibilidad de recorrer y conocer el museo a través de sus propios medios.

Los visitantes que llegaban al museo sólo contaban con la ayuda de los guías para poder enriquecer sus conocimientos sobre los distintos objetos y, en algunos eventos especiales, podía observarse el funcionamiento de los mismos, cuando el personal del museo los sacaba de las vitrinas para exponerlos en una mesa central. Por otra parte, el museo no disponía de ningún sistema de información que permitiera administrar la información relacionada a los objetos en exposición.

En este sentido, era de suma importancia para el museo disponer de un sistema, que no sólo le permita administrar la información de los objetos, sino también que le brinde al visitante la posibilidad de recorrerlo en forma autónoma, sin la necesidad de

solicitar un guía, y a su vez, de poder tener diferentes vistas de los instrumentos aún estando dentro de sus vitrinas.

3 La Mejora Tecnológica Propuesta

Al observar el desarrollo tecnológico que ha experimentado la humanidad desde mediados del siglo XX hasta la actualidad, hace pensar que más que un avance, se ha producido una verdadera revolución. Mantenerse conectado a Internet y manipular información digital desde cualquier parte y en cualquier momento, se han convertido en necesidades básicas para la sociedad de estos tiempos. Y, conforman un campo propicio para innovar tecnológicamente en este sentido, en una entidad abierta al público y de interés general, como es el Museo de Física.

En un principio se propuso aprovechar las potencialidades que ofrecen los dispositivos móviles como ser su tamaño reducido, portabilidad, incorporación de cámaras, sensores incluidos como GPS (Global Positional System), brújula digital, acelerómetros y giroscopios, importante capacidad de procesamiento gráfico y almacenamiento de datos [4], para dar lugar al desarrollo de una aplicación de realidad aumentada (RA) [5]. Pero, esto no fue factible de realizar por las condiciones físicas que presentaba el Museo de Física. Debido a que el museo se encuentra en un único habitáculo sin ventanas, donde los instrumentos se encuentran apilados, con poca luz, detrás de vitrinas cerradas, donde se produce mucho reflejo a través del vidrio de las mismas. Las vitrinas tienen la misma ubicación tanto en el primer piso como en la planta baja y son recorridas a través de pasillos laterales.

Estas condiciones y factores en el interior del lugar, hicieron que no fuera posible la utilización del sensor GPS, ni reconocimiento de imágenes a través de la cámara del celular, ya que se lograba poca identificación de los instrumentos. La realidad aumentada fue descartada.

Por tal motivo, la propuesta definitiva y consensuada de trabajo conjunto entre el Museo de Física y la Facultad de Informática, ambas unidades pertenecientes a la UNLP, fue el aplicar aumento de la información de los dispositivos e instrumentos expuestos en el museo, mediante código QR, para gestionar distintos tipos de contenidos, en un entorno científico, educativo y cultural, en pos de ofrecer al visitante una experiencia más personalizada e innovativa.

Específicamente, se incorporaron los dispositivos móviles al museo, como herramientas para optimizar la captación y observación de los instrumentos en exhibición, aumentando la información percibida por el visitante, con elementos adicionales sintetizados y presentados en distintos formatos multimedia.

Para ello, se desarrolló una aplicación móvil que permite al visitante investigar sobre el museo y acceder a información adicional que se dispone de los objetos que se exponen en él, pudiendo observar más allá del objeto real, videos del mismo en funcionamiento, explicaciones textuales, visuales y auditivas, acceso a sitios Web que lo referencian, juegos educativos sobre el tema y encuestas relacionadas para incentivar la participación del usuario.

El visitante puede interactuar con la aplicación de una manera activa, construyendo sus propios recorridos del museo, personalizados y auto gestionados, como también configurando por sí mismo, el nivel de información que desea obtener de los objetos de interés y el formato preferido para observarlos.

El personal del museo además de dirigir las visitas guiadas tradicionales, tendrá un nuevo rol como ser el de observar el comportamiento del usuario frente a estos nuevos recursos tecnológicos brindados, acompañar al mismo en sus recorridos personales, supervisar y/o evacuar consultas. Además, administrar la información digitalizada en sus distintos formatos.

Para esto último, también se desarrolló una aplicación Web administrativa, que permite al personal del museo administrar toda la información adicional, relacionada a los objetos en exposición, como textos, páginas, imágenes, audios, videos y juegos. Incluye un generador de códigos QR Toda la información está alojada en un servidor de base de datos que no poseían, la cual fue diseñada especialmente en este proyecto y que constituye también una fuente de información centralizada, para nutrir el sitio Web oficial del museo con estos nuevos elementos multimedia.

4 El Código QR

En pos de implementar la propuesta de mejora tecnológica enriqueciendo y aumentando la información del objeto o instrumento de exhibición que el usuario está percibiendo mediante el uso de dispositivos móviles, se trabajó sobre el concepto de código QR o marcadores.

Un marcador, denominado en inglés *fiducial markers*, es una imagen 2D impresa con un formato específico reconocido por la aplicación que se desarrolla [6].

En la Fig.1 pueden verse diferentes tipos de marcadores de los que existen lectores de código abierto.

(a) (b) (c) (d)

Fig.1. Diferentes tipos de marcadores: template(a), ID-Marker (b), DataMatrix(c), QRCode (d*)*

Los marcadores *template* (Fig.1 a) son muy conocidos dado que son los utilizados por la librería de realidad aumentada ARToolkit [7], que fue la primera librería que popularizó las aplicaciones de realidad aumentada. El formato es un cuadrado negro y dentro del mismo se encuentra un cuadrado blanco que tiene en su interior, una imagen asimétrica en negro.

Los marcadores *ID-marker*, que se observan en la Fig.1 (b), fueron los marcadores utilizados en este proyecto. Estos marcadores codifican un número de 9-bits (hasta 512 diferentes) en un patrón de 6 x 6, repitiendo los 9 bits 4 veces completando los 36 bits. Una variante de estos marcadores son los denominados BCH (Bose, Ray-Chaudhuri, Hocquenghem), los cuáles son más robustos que los anteriormente

descritos, ya que usa un algoritmo avanzado de chequeos de redundancia cíclica (CRC) que permite restaurar marcadores dañados. Se incrementa el número de marcadores disponibles a un total de 4096.

Los marcadores *DataMatrix* y *QRCode* (Fig.1 c y d) no fueron diseñados específicamente aplicaciones como las de realidad aumentada, sino que su propósito inicial es codificar una serie de caracteres ASCII. Uno de los usos más comunes es la codificación de una dirección Web o URL de forma que una aplicación al leerlos y decodificarlos, pueda derivar al sitio web codificado. Por esto, su uso principal se asocia a los hipervínculos.

Mientras que los *DataMatrix* pueden almacenar hasta 2335 caracteres, los *QR Code* almacenan 4296 caracteres. Una diferencia entre ellos radica que el *QRCode* incluye además símbolos japoneses. Ambos códigos son abiertos y pueden descargarse de forma gratuita aplicaciones lectoras para los teléfonos celulares del mercado [8].

5 Descripción de la Aplicación Móvil para el Museo de Física

La aplicación móvil que se desarrolló para el Museo de Física, fue implementada en Java sobre el framework Spring bajo el sistema operativo Android [9].

Una vez que el usuario ingresa al museo, puede hacer uso de esta aplicación móvil, descargada previamente desde el sitio oficial del museo. El aplicativo posee una interfaz del usuario que respeta los estándares de diseño para móviles [10].

El personal del museo tiene a disposición varios dispositivos móviles con la aplicación ya cargada por si algún usuario quiere experimentar ese nuevo modo de recorrido por el museo y no tiene los recursos tecnológicos para hacerlo.

La primera pantalla que se observa al iniciar el sistema, presenta cuatro actividades principales como se muestra en la Fig.2. Las actividades son: *Sobre el museo*, donde se presenta su historia y mapa; la actividad *Explora*, que permite hacer uso de la identificación de objetos con código QR y aplicar el aumento de la información; la actividad *Encuesta*, para el feedback del usuario con sus opiniones, sugerencias, y grado de conocimiento adquirido; y finalmente, la actividad *Ayuda* que permite guiar al usuario sobre el uso del sistema.

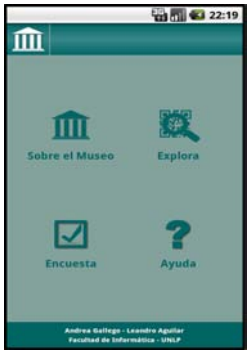


Fig.2. Pantalla principal

La actividad *Sobre el Museo* (Fig.3), tiene el objetivo de mostrarle al usuario información básica acerca del mismo y un mapa de las instalaciones. La información que se visualiza aquí es la misma que la del sitio Web tradicional. Se obtiene el contenido directamente del sitio oficial mediante una tarea asincrónica que logra la comunicación con el servidor y el pasaje de la información.

Fig.3. Actividad "Sobre el Museo"

La actividad *Explora* es la componente funcional principal de la aplicación móvil. Cuando el usuario selecciona esta opción del menú, se ejecutará la tarea dedicada a la lectura del código QR. Una vez que detectó y leyó el código, se procesa el resultado que se recibe de la lectura, la cual representa el identificador del objeto sobre el cual está interactuando el usuario.

A partir de aquí, el usuario tendrá la posibilidad de observar información adicional de dicho objeto tanto en formato texto, como también en audio (Fig.4).

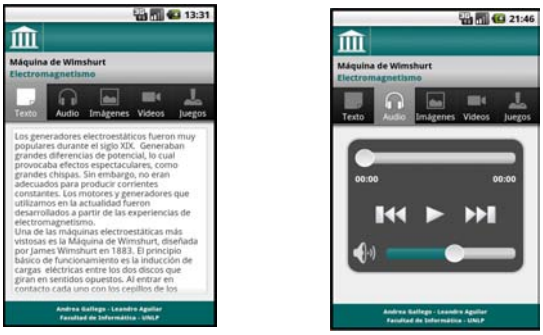


Fig.4. Actividad *Explora*, pestañas *Texto* y *Audio*

También. el usuario podrá visualizar contenido adicional mediante imágenes y videos relacionados al objeto, que aumentan y enriquecen la información transmitida al usuario. Esto se muestra en la Fig. 5, donde se observan las pestañas de la aplicación destinadas para esto.

A través de esta sección, el usuario puede visualizar un conjunto de imágenes asociadas al objeto. En principio, se puede ver el listado de imágenes con su nombre y una breve descripción de la misma. Cuando el usuario selecciona una de ellas, la

imagen entera se visualizará en tamaño completo y el usuario podrá navegar hacia la izquierda o derecha entre las imágenes disponibles, sin necesidad de volver al listado.

En este sector se utiliza un escuchador para los eventos táctiles que se produzcan en la pantalla, así se puede interpretar los gestos que el usuario realiza con el dedo. Si hace un arrastre, el sistema lo interpreta como movimientos entre imágenes disponibles, en cambio cuando toca la pantalla, el sistema muestra y oculta la descripción de la imagen activa.

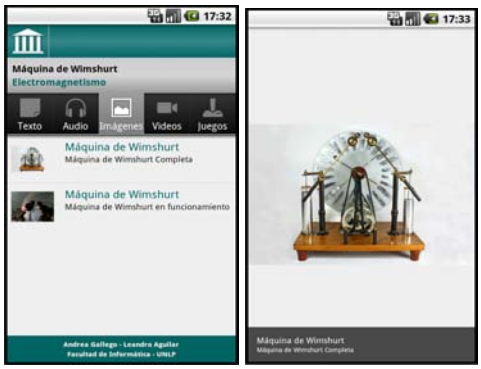


Fig.5. Actividad "Explora", pestaña Imágenes

Un aspecto de la aplicación que permite agilizar la interacción, es que la imagen se guarda en una memoria caché para tenerla disponible en caso que el usuario quiera girar la pantalla para verla en distintas posiciones. De esta forma, se evita comunicarse nuevamente con el servidor.

En el caso de seleccionar videos, se muestra la pantalla visualizada en la Fig.6. En esta sección el usuario podrá ver distintos videos relacionados al objeto. Se debe seleccionar uno de los videos listados y se lo puede visualizar en pantalla completa.

La información del video que se visualiza en la Fig.6 y que el usuario seleccionó previamente, tiene un identificador de un video de YouTube [11], como ser la url http://www.youtube.com/watch?v=Zilvl9tS0Og el id es Zilvl9tS0Og.

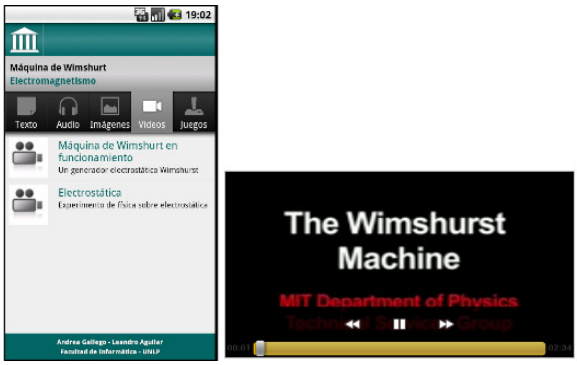


Fig.6. Actividad *Explora*, pestaña *Videos*

Entonces, para poder reproducir este video, se debió acceder al archivo xml con la información que se corresponde a ese identificador [12]. Para ello, YouTube provee

la url https://gdata.youtube.com/feeds/api/videos/, que es su protocolo del API de datos. Dentro de este archivo xml encontramos la url del video con formato 3gp el cual soporta la aplicación.

En el caso de la pestaña *Juegos*, se muestra al usuario una pantalla con dos tipos de juegos educativos específicos para el objeto que se está visualizando, el *Trivia* y *Unir con Flechas* (Fig.7). De esta manera, se puede evaluar lo que el usuario aprendió del objeto que estuvo observando, como también analizar la efectividad de la aplicación como mecanismo comunicacional y educativo.

Fig.7. Actividad *Explora*, pestaña *Juegos*

Los contenidos de los juegos *Trivia* (Fig.8) y *Unir con Flechas* (Fig.9) son dinámicos y aleatorios, ya que dependen del objeto y de las opciones cargadas en una base de datos del servidor.

Fig.8. Actividad *Explora*, pestaña *Juegos*, *Trivia*

El juego *Trivia* presenta otros aspectos dinámicos, como ser el tipo de las preguntas que se generan en el momento, las cuáles pueden ser de selección múltiple o del estilo Sí/No, y el tipo de respuesta, con una opción correcta, varias o ninguna. Es por este motivo, que los distintos componentes utilizados para representar las preguntas y respuestas del juego, se crean dinámicamente al iniciar la actividad.

Una vez que el usuario responde todo, se valida la cantidad de respuestas correctas que realizó y se muestra los resultados al usuario a través de un cuadro de diálogo. Estos resultados se registran en el servidor para poder ser analizados posteriormente.

El desarrollo del juego *Unir con Flechas*, consiste en relacionar un concepto de la columna izquierda con uno de la columna derecha. Para ello, el usuario debe

seleccionar un concepto de alguna de las columnas para activar la flecha y arrastrar la misma hacia el concepto de la otra columna que el usuario considera que se corresponde. El juego termina cuando logre unir todos los conceptos (Fig.9).

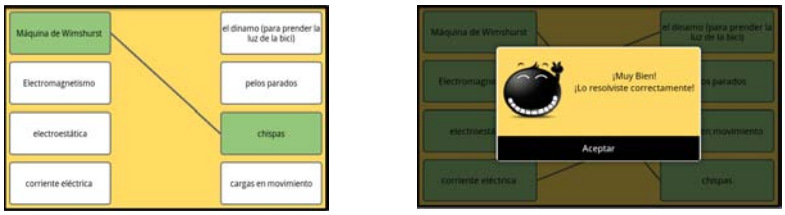


Fig.9. Actividad *Explora*, pestaña *Juegos*, *Unir con Flechas*

Cuando el usuario une todos los conceptos, se le muestra un diálogo indicando la cantidad de asociaciones correctas. También, la jugada se guarda en el servidor para analizar posteriormente el grado de comprensión de los conceptos relacionados con el objeto que el usuario adquirió con la experiencia.

También, se tiene la actividad *Encuesta* (Fig.10) que el usuario debe responder. Luego, el personal del museo tendrá acceso a los resultados a través del sitio Web de administración, como sucede con el tema de los juegos, lo que les permitirá sacar distintas conclusiones con respecto a la visita que realizó el usuario.

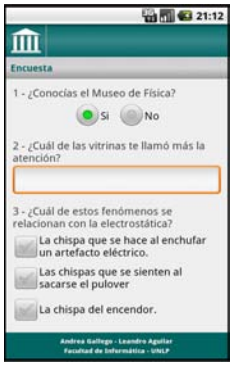


Fig.10. Actividad *Encuesta*

La carga de la encuesta se realiza dinámicamente al igual que se hace con *Trivia*, pero aquí se permiten además, respuestas abiertas para que el usuario escriba con el teclado virtual que se activa automáticamente. Otra diferencia que se tiene, es que aquí no hay corrección, sino que simplemente, se envían los resultados al servidor para que puedan ser almacenados y consultados posteriormente.

6 Conclusiones

El Museo de Física es un establecimiento que alberga más de 2000 instrumentos de Física en un ambiente único y cerrado, para el cual se desarrolló una aplicación móvil

para el aumento de la información exhibida basada en código QR. El mismo tuvo como objetivo ofrecer entornos más verosímiles que permitan al visitante una mayor inmersión, generando la posibilidad de que interactúen directamente con los objetos expuestos de una forma atractiva y a la vez didáctica. El mundo virtual y sintético lleva al museo, lo que el mundo real no puede, gracias a la capacidad de insertar objetos virtuales en un espacio real.

Mediante una aplicación móvil sencilla, fue posible aumentar la información transmitida de la colección de los instrumentos que se pueden observar en el museo, ofreciendo distintos formatos de contenidos como videos, imágenes, texto y audio, y también de actividades como explorar, jugar, participar.

Este concepto de participación activa ofrecida a través de la exploración con la aplicación, las encuestas y los juegos, permitió manifestar la interactividad en tiempo real. No sólo se pretendió mostrar información aumentada y complementaria de los instrumentos en exposición, sino que se intentó producir una retroalimentación de la experiencia con el participante.

Teniendo en cuenta la afinidad de los jóvenes con las nuevas tecnologías, el personal del museo espera que esta nueva interfaz sirva para acercar a los estudiantes a las temáticas que abarca el museo, habitualmente fuera de los intereses de la mayoría de los niños y adolescentes. Por otra parte, el número siempre creciente de visitantes a la página de internet del Museo los obliga a mejorar y actualizar constantemente los contenidos, y este proyecto permitió adicionar y centralizar nuevos materiales digitalizados.

Referencias

1. Von Reichenbach, M.C; Cabana, M.F: El Museo de Física como vínculo entre la extensión y docencia formal. Universidad Nacional de La Plata- CONICET.
2. Ajay R. Mishra. Cellular Technologies for Emerging Markets: 2G, 3G and Beyond. John Wiley & Sons, Ltd. 2012
3. Sitio oficial del Museo de Física de la Fac. de Cs. Exactas, UNLP. museo.fisica.unlp.edu.ar/
4. Zhou, F.; Dhu, H. and Billinghurst, M. (2008) "Trends in Augmented Reality Tracking, Interaction and Display: A Review of Ten Years of ISMAR" Proceedings of 7th IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR 2008), pp 193-202.
5. Azuma, R.; Baillot, Y.; Behringer, R.; Feiner, S.; MacIntyre, B. (2001) "Recent Advances in Augmented Reality". IEEE Computer Graphics and Applications, 21(6), 34-47.
6. Denso Wave Corp. Documentation of the QRCodes <http://www.denso-wave.com/qrcode/index-e.html>
7. ARToolkitPlus [en línea] <http://handheldar.icg.tugraz.at/artoolkitplus.php>
8. Saquele Jugo a los Código QR. Conozca el potencial de escanear Código QR en smartphones. Fernando Toledo.
9. Gargenta Marko. Learning Android. 1° Edición, O'Reilly Media, Inc. 2011.
10. Designing Effective User Interfaces for Wireless Devices. Reza B'Far with Roger Richards and Stephen Ditlinger.
11. Guía para desarrolladores. Protocolo API de datos de YouTube. developers.google.com
12. XML - Edición 2012, de Miguel Ángel Acera, Editorial Anaya

Implementation of a 3D Virtual Environment at the National University of the North West of the Province of Buenos Aires

Hugo Ramón[1], Claudia Russo[1], Leonardo Esnaola[1], Nicolás Alonso[1], Maximiliano Fochi[1], Franco Padovani[1]

[1] Institute of Research and Technology Transfer (ITT), University of the North West of the Province of Buenos Aires (UNNOBA), Junín, Buenos Aires, Argentina

{hugoramon, crusso}@unnoba.edu.ar, {leonardo.esnaola, nicolas.alonso, maximiliano.fochi, franco.padovani}@nexo.unnoba.edu.ar

Abstract. The increasing incorporation of ICTs in education is generating a series of changes and transformations in the ways we represent and carry out teaching and learning processes. These changes can be seen in the traditional environments of formal education but also in the arrival of new educational environments totally or partially based on ITCs. This document discusses the design, creation and coordination of a 3D virtual teaching and learning environment for UNNOBA. With that aim, the definition and deployment of a 3D virtual environment [3DVE] focuses not only on technological and pedagogical aspects but also on a transition methodology to improve the quality of teaching methods and techniques.

Key words. 3D virtual environment, PACIE, OpenSim, ICTs.

1 Introduction

In the past few years, we have seen that education has been transformed as a result of the advent and introduction of new technologies, especially since the introduction of LMS (Learning Management Systems). These systems favored the growth of distance learning and helped to improve teaching and learning processes. Despite these advances, LMS providers compete to place their tool as the central element of e-learning [1] leaving aside what should be the main objective i.e., allowing students to assume more responsibility for their learning processes.

In the traditional approach of current LMS, teachers design courses and activities with a student-centered vision but this does not allow students to set their own goals and terms, manage their work, assess their use of time and work in collaboration with their classmates [2].

The impact of these new technologies has a lot to do with the advent of a new generation, born with technology at their fingertips, known as *digital natives* [3], who prefer to learn in an ICT tool-based environment. This is a challenge for traditional teachers who need to generate new methods or adapt their old ones so as to exploit students potential through new technologies [4].

As a result, we believe that, in order to design, create and coordinate a 3D virtual teaching and learning environment for UNNOBA, we should not focus on technical or pedagogical aspects only. We need to consider a transition methodology to improve teaching methods and techniques. For future implementation purposes, we will base part of the design and creation of our 3DVE on the PACIE method (Spanish acronym for presence, scope, training, interaction and e-learning) developed by Pedro Camacho [5].

2 Key concepts and aspects of the PACIE method

PACIE is a method to use ICTs as support for learning and self-learning processes that enhances a pedagogical framework of real education. This method differentiates three aspects that should be taken into account in the transition process for a 3DVE, as follows:

- The image to be given to our 3DVE
- Information management and organization within the 3DVE

- Virtual education management, organization and administration.

Our 3DVE proposal will focus on these three fundamental aspects. Other aspects will be open to inquiry until we have carried out different experiences after the 3DVE start-up [6].

2.1 3DVE image

The environment needs to be friendly and attractive. The following should be taken into consideration:

- Using the same typography and text size for titles.
- Using the same font for information.
- Highlighting relevant information by using different fonts and colors.
- Using the same size for all images.
- Using links with images to facilitate access to different items of the University.
- Including attractive web 2.0 resources such as animations, videos and other.
- Creating the need to discover new and appealing ways in 3DVE to motivate students to continue exploring and using it.
- Having and keeping an identity. This means making students feel that each virtual classroom is part of a single virtual education center [7].

2.2 Information Management and Organization

Now we should consider the importance of managing and organizing information inside our virtual classroom.

Based on the idea that to let students learn, the aims should be clear, we decided to use SBS (*Standards, Benchmarks and Skills*) in the proposed model. By using them, we can determine what skills the students need to acquire, develop or improve when they complete the educational process and achieve certain aims or products by meeting certain requirements and being ready to face new situations. All these skills are grouped into patterns or academic benchmarks with similar characteristics. All of them have been customized for our subject in an exclusive way. Once we have the academic benchmarks, we aim at certain standards that need to be established by the related academic department [8] [9].

2.3 Virtual Education Management, Organization and Administration

Following the PACIE method, to implement the proposed 3DVE, we believe that it is very important to have a specialized area of distance learning to manage and organize all the aspects related to virtual education, with decision-making powers over the 3DVE.

As stated in the PACIE method [10], the minimum human resources requirements include a teaching expert, an IT expert and a social communication expert. This allows us to divide tasks in a better way.

- The social communication expert should manage communication of virtual learning processes as well as any general aspect related to the 3DVE and the virtual classrooms. This person should be the link between the teaching expert and the IT expert.
- The IT expert should provide support in all aspects related to technology and the teaching and learning environment by providing solutions for the 3DVE to work properly.
- The teaching expert should be in charge of implementing learning methods suitable for the 3DVE and thus generate better learning techniques within the 3DVE and a teaching method especially designed for this specific 3DVE.

3 Identification of the main aspects of the proposed 3D virtual environment

As a result of the research process, a series of issues and specific solutions for the model have been identified:

- *Administrative aspect*: A 3DVE is a system that integrates different solutions but its main objective is to provide adequate support to the teaching and learning activities of the main players: teachers and students alike. They need strong guidance to be provided by people who are highly trained and supported by well-defined processes.
- *Information management aspect*: We should put special emphasis on information management and organization, analyze what is done with such information and how to use it to generate learning. We can manage and organize information using current standards or methodologies and using SBS is one of them.
- *Educational aspect*: Teachers need to break with the traditional teaching scheme, they should guide students toward self-learning, make the environment a door to contents students find interesting so that they will continue exploring them and learning autonomously. This is done by developing adequate contents for the environment and guiding students through it [11].
- *Technical aspect*: A good 3DVE should be accompanied by people who are trained to use it and can coach those who have not learned how to use it.
- *Evolutionary aspect*: Any virtual system should take into consideration a potential evolution if it aims at being effective in changing times such as the present. Control and update process are necessary to adapt the system to new tools and create a virtuous circle to have it permanently optimized.
- *Graphic aspect*: Nowadays, the design and visual aspect of the 3DVE needs to be attractive to students, otherwise, they are likely not to access it very frequently and, as a result, they will not make use of its advantages.
- *Functional aspect*: Necessary tasks such as the management of virtual classrooms: their creation, enrolment process, teacher appointments and role allocation, etc., should be standardized and carried out by specific personnel. On the one hand, because these tasks should not be part of the teacher's responsibilities, the teacher is a user and not an administrator of the virtual environment but on the other hand, because the performance of these tasks needs to be uniform, since a diverse environment looks messy.

4 Choosing a tool to implement the proposed 3DVE

To implement the proposed 3DEV, we needed to go through a selection process for the IT tool to be used as a server for the 3D applications. One of the requirements in the process was that the tool needed to be open source. *OpenSim*[1] and *OpenWonderland* were among the 3D application servers we surveyed and researched into.

OpenSim and OpenWonderland are 3D servers that allow virtual environments (also known as virtual worlds) to be created. These environments can be accessed from a wide range of viewers, also known as clients. *Singularity*[2] and *Imprudence*[3]. are two of the most popular OpenSim viewers. In the case of OpenWonderland, the web server acts directly as a viewer. No additional software to act as a client and interact with the 3DVE is needed.

When we chose the tool, we tried to meet the requirement that the 3DVE could connect to the virtual teaching and learning environment (EVEA) currently in use at UNNOBA, known as UNNOBA Virtual [12]. OpenSim had existing modules that allowed us to meet this requirement. This, and its excellent graphic quality, the possibility of making synchronic voice communication, its various repositories and documentation available made OpenSim the best option to be used as a 3D application server.

[1] OpenSim official site, http://opensimulator.org/wiki/Main_Page
[2] Imprudence official site, http://wiki.kokuaviewer.org/wiki/Imprudence:Downloads
[3] Singularity official site, http://www.singularityviewer.org

After choosing the tool, we continued working on the installation stage. To do so, we downloaded the latest version available on the official website (we are currently working on the 0.7.5 version) and we consulted the official documents available on the site so as to install it correctly.

After completing the installation, we started working on modeling a virtual world. The first step was to create a conference room for UNNOBA. To do, we downloaded a base building and free objects from different websites, such as *FleepGrid*[4], *OpenSim-Creations*[5] and *Zadaroo*[6], among others, to give it the current look and feel through different modeling and object options associated with the viewers, as shown in figures 1 and 2.

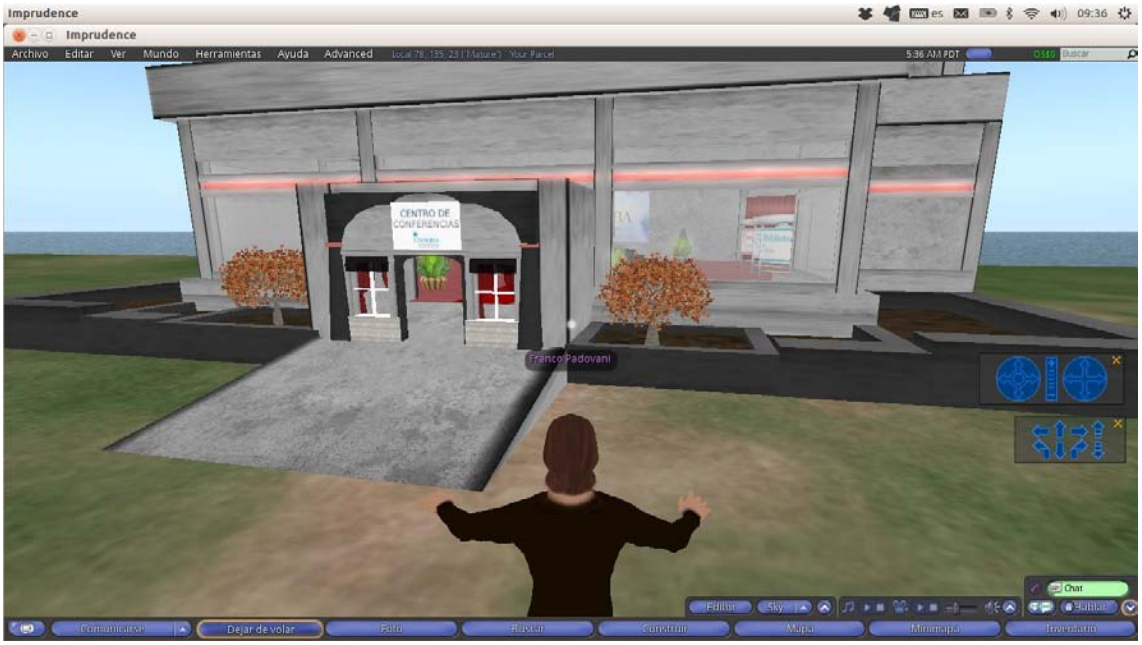


Figure 1. Conference room outside view.

Figure 2. Conference room inside view.

[4] FleepGrid official site, http://fleepgrid.com/store
[5] OpenSim-Creations official site, http://opensim-creations.com
[6] Zadaroo official site, http://zadaroo.com

As second stage of the virtual world modeling, we worked on the idea of recreating one of UNNOBA's buildings. For such purposes, we had already obtained a blueprint of the building.

During this stage, research was conducted on the ways in which OpenSim would allow buildings to be created in its virtual worlds. Therefore, three ways of modeling and building were assessed using OpenSim. These are:

- Building them with a 3D modeling program and saving them in an extensible markup language, in the .XML format, compatible with OpenSim.
- Building them in a 3D modeling program and saving them with the .DAE format, also known as mesh, as this format is also compatible with OpenSim imports.
- Building them manually, object by object, using the viewer.

Out of these three options, we opted for a 3D modeling program that allows exporting .DAE files, because we had experience using programs with this format, such as *Blender* y *Google Sketchup*[7], and, the *Singularity* viewer, that is the one used by the 3DVE, supported .DAE files.

We did not choose to create models manually because of the complexity of the task of creating buildings using primitive objects such as the spheres, squares and cylinders available on the viewers. The option of creating models and saving them on an extensible markup language was discarded because the tool we had, *SolidWorks*[8] , presented us with several inconveniences when importing the model into the 3DVE during the technical assessment stage.

Using the selected option, we managed to import, inside the 3DVE, a model based on a blueprint of one of UNNOBA's buildings. This is shown in figures 3 and 4.

Figure 3. UNNOBA's building constructed from Google Sketchup already imported into OpenSim.

[7] Google Sketchup official site, http://sketchup.google.es/index.html
[8] SolidWorks official site, http://www.solidworks.com

Figure 4. Library UNNOBA's building constructed from Google Sketchup already imported into OpenSim.

5 Integration and technical assessment of the proposed 3D virtual environment

Once the virtual world mentioned above was created, we started its technical assessment.

One of the aspects we analyzed and assessed was real-time voice interaction. This was one of the requirements to be met. In these communication modes i.e., open voice interaction and close voice interaction, for public and private communication respectively, we detected that open voice communication can be configured so as to listen within a set distance, which would allow splitting working stations and allow only the avatars in the area to listen and be listened. "Avatar" is the name given to the virtual representations of the human body of the user in the environment.

Another aspect to be taken into consideration was the possibility of integrating the 3DVE into UNNOBA's teaching and learning environment, UNNOBA Virtual [12]. To do so, we used a module known as *Sloodle*[9] that allows integrating a *Moodle*[10] based LMS, the one used in UNNOBA Virtual, and OpenSim, the tool selected for the implementation of the proposed 3DVE. By using Sloodle, any component created on UNNOBA Virtual could be accessed from the 3DVE. The main technical assessments and trials conducted in connection with this integration were as follows:

- Access using existing access credentials from the the 3DVE to the virtual teaching and learning environment used at UNNOBA.
- Creating presentations in a virtual classroom of the environment and being able to access them from the 3DVE.
- Taking exams or assessments in a virtual classroom used at UNNOBA allowing students to complete them from the 3DVE, thus enriching the virtual assessment experience.
- Integrating a synchronic communication tool, as the chat used at UNNOBA's virtual environment, with the 3DVE, which gives the possibility of starting a synchronic written and verbal communication among the participants in a virtual classroom.

[9] Sloodle official site, http://www.sloodle.org
[10] Moodle official site, https://moodle.org

6 Conclusions and future work

One of the aims we seek to attain by using the 3DVE is to introduce and generate innovative strategies to meet the need to improve the teaching and learning process, allowing students to be involved in their academic education and favoring open and dynamic communication.

We believe that although we are still on an assessment and continuous improvement stage, the proposed 3DVE, thanks to its real-time interaction capacity and its "face-to-face" feel, adds a social dimension to the teaching and learning environment similar to the one of "face-to-face" education which makes our distance learning proposal richer and more dynamic.

However, we are aware of the fact that to improve this process, having and implementing a 3DVE is not enough. Having a transition methodology to improve the quality of teaching methods and techniques is not enough either. The design of educational activities and the generation of specific content for the 3DVE is an emerging perspective in the practice and research of the e-learning community.

Therefore, we should focus on generating and designing specific content for this kind of environments and, as a result, in the future we will attempt to design specific content for the 3DVE to be used at UNNOBA.

In connection with our future work, with the integration of the teaching and learning environment used at UNNOBA and the 3DVE, we aim at creating a collaborative learning environment and designing specific content and activities suitable for both environments and also at assessing, measuring and comparing the impact of the introduction of a 3DVE into the different teaching modes at UNNOBA.

References

1. Alan Palme. Enhancing learning and teaching through the use of technology: a revised approach to HEFCE's strategy for e-learning. HEFCE. United Kingdom, 2009
2. Master's Degree in Instructional Design by José Luis Córica, http://cvonline.uaeh.edu.mx/Cursos/Maestria/MGIEMV/DisenoProgramasEV12/materiales/Unidad%204/Cap4_DisenoInstruccional_U4_MGIEV001.pdf
3. Prensky, M. (2001). Digital natives, digital immigrants. On the Horizon , 9 (5), 1-6.
4. Journal *La educación y la Virtualidad Editorial: Grupo Dseta Editora: Francia Tovar Romero* (Education and the virtual world Journal), http://www.youblisher.com/p/173589-Please-Add-a-Title-La-Educacion-y-la-Virtualidad
5. The PACIE method, essay published by Luis Oñate for FATLA, 2009, http://iuetaebvirtual.wikispaces.com/file/view/22234756-La-Metodologia-Pacie.pdf
6. Essay *Plataformas de educación a distancia*, ("Distance learning platforms") Rambo Alice, http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Educacion_Distancia_Alice_2009.pdf
7. PACIE methodology applied to virtual environments, http://pacie-en-muves.wikispaces.com/home
8. Presentation *Aplicación PACIE en los Estándares Académicos de la Educación Virtual* (Applying PACIE to academic standards in virtual learning) by Deizi Carolina De Jesús Lobo, http://pacieeducavirtual.jimdo.com/aplicaci%C3%B3n-de-pacie-en-los-est%C3%A1ndares-acad%C3%A9micos-de-la-educaci%C3%B3n-virtual-aula-virtual-moodle/
9. Educational experience with 3D environments, UPEL (Universidad Pedagógica Experimental Libertador), http://www.ugr.es/~sevimeco/revistaeticanet/numero10/Articulos/Formato/articulo5.pdf
10. Module implementation based on the PACIE method, 2011, http://www.moodlemoot.org.uy/moodlemoot_2011/moodlemoot/moodlemootuy2011_submission_25.pdf
11. Essay *Entornos Virtuales 3D, Alternativa Pedagógica para el Fomento del Aprendizaje Colaborativo y Gestión del Conocimiento* (3D virtual environments, an educational alternative to foster collaborative learning and knowledge management), http://www.scielo.cl/scielo.php?pid=s0718-50062011000200006&script=sci_arttext
12. UNNOBA's virtual teaching and learning environment, http://virtual.unnoba.edu.ar/

Web Authoring Tool and Repository for Learning Objects

Lucas Ferrari da Costa, Maximiliano Reidel, Vinícius de Carli,
Júlia Marques Carvalho da Silva,

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) – Campus Bento Gonçalves, Av. Osvaldo Aranha 540, 95700-000 Bento Gonçalves, Brazil
{lukas.ferrari95, vinidcali}@gmail.com, {maximiliano.reidel, julia.silva}@bento.ifrs.edu.br

Abstract. This article purposes to present an authoring tool for learning objects (LOs) with metadata filling in the OBAA standard and SCORM packaging generated in scientific initiation research. The software, in early stage of development, is noted to be free and easy to use, programmed for the Web platform pointing towards features to enhance the user experience. The application is internationalized and has a repository function, which stores the LOs in a database accessed through the search feature. Its origin took place on the lack of tools making use of the characteristics applied to this one and its goal is to simplify, to the teacher, the creation of teaching materials, as well as encourage students to invent their own content. Therefore, a study was conducted of LOs, their metadata, specifications and related tools as well as of the technologies needed to create the system, such as concepts, languages and programming standards.

Keywords: Learning Objects, OBAA, SCORM, e-learning.

1 Introduction

According to Wiley [1], Learning Objects (LOs) are digital educational content, whose basic feature is their ability to be used on other occasions. Thus, LOs are made up of texts, images or videos, for example, along with their metadata, which, together, form a reusable LO. The metadata are essential to the educational material since the repositories use them to catalog and store objects. To facilitate cataloging, several metadata specifications have been developed, such as IEEE LOM (international reference standard), CanCore (Canadian standard) and OBAA (Brazilian standard).

To simplify the generation of LOs and their metadata, there have been developed some authoring tools. However, most of them present themselves complex, require computer skills or additional installation to be used, or, still, demand payment. Even so, there are few which allow the making of objects in the OBAA standard or make use of SCORM during the files packaging.

The present research, conducted focusing on the relevant basic concepts and on the current situation of the subject, as well as on the technology needed to create the software, has been characterized as preparation for the construction phase, which

generated a tool - at the time in early stage of development - of simple and easy procedure, that, besides standardizing the metadata description and LO packaging, requires only a web browser with Internet access to work. It was also sought to adapt the tool to different devices and internationalize it. It is expected, with its use, to facilitate the teacher's task to create educational content, encourage students to invent their own materials and disseminate the OBAA model.

This document is divided into four chapters. First there is an introductory text about the covered issues, followed by the Theoretical References appropriate to the research phase (split into Methodology Used, OBAA standard, SCORM standard and Authoring Tools). In third takes place the presentation of the proposed tool, with details about its project and development as well as examples of its current operating condition. Finally, it is concluded the article and elucidated how the project will be continued.

2 Theoretical Reference

2.1 Methodology Used

Until the present moment, this work is based on qualitative research because it used bibliographic research approach in order to understand the concepts to be smeared on the tool. Next, this study turns into applied research since it aims its use in practice, involving future instructional designers as subjects.

Initial research marks itself with online studies, in websites as well as in academic papers, aiming to understand the fundamental concepts connected to the project subject, such as the study of the PHP language and the theories and applications of learning objects (LOs) and their metadata. After, there has been a reading of Wiley's [1] and Silva's [2] works, which are widely cited in papers of their areas, in order to supplement the initial concepts notion. It has also been sought to understand the current situation of the tools and technologies available for the creation and maintenance of LOs.

Following the initial research stage, it has been run a study of the languages and tools compulsory to the software creation (emphasis on PHP, JavaScript, CSS, HTML, Database/MySQL), and, from the model proposed by the mentor teacher, it was then given start to the development. Additionally, it has been adopted the OBAA standard of metadata descripting through XML files and the SCORM standard for file packaging.

The following research steps aim to verify using it in the potential scenarios: descripting a LO, packing a LO, searching for a LO on the repository, etc. For that, it will be developed schedules which, when applied, it will be adopted non-participant observation, in order to register the interaction with the tool. Furthermore, it is possible to include a log record which will indicate how and when the users used the system, supplying quantitative and qualitative data.

2.2 OBAA

Knowing that a LO is any educational media that can be reused, OBAA has emerged as the Brazilian proposal for the educational and technical description of these objects through metadata, in order to facilitate their retrieval and access. The model, whose acronym translates to "Agent-Based Learning Objects", was developed by UFRGS in partnership with UNISINOS, aiming to specify standards for technical and functional requirements in response to interoperability problems of some digital contents that, in short, were settled by the XML syntax typical of this standard [2].

OBAA was established in the IEEE LOM, an international reference standard extensively used which allowed some changes that were considered necessary for the creation of OBAA. Consequently, OBAA aggregated the LOM categories, applying modifications and adding new items [3].

2.3 SCORM

SCORM (Sharable Content Object Reference Model) is a collection of specifications and standards which define the interrelationship of content objects and data models so that the objects are sharable on systems that follow this model [4]. This specification promotes reusability, accessibility, durability and interoperability of learning content, and facilitates migration between different learning management systems. SCORM is responsibility of Advanced Distributed Learning and was originally released in 2000, with the latest version being the SCORM 4th Edition, of 2009.

Tarouco and Dutra [5] explain that resources (also called educational media) are the smallest physical units within the SCORM material. Examples include web pages, images, videos and flash applications. Its main feature is to be reusable, and for this to be possible, it is used the asset's metadata. SCOs (Sharable Content Objects) are the sets of resources which represent the smallest logical units of the material: they can represent a class, a topic or a module in a course. SCOs cannot communicate among themselves because they are independent.

2.3 Authoring Tools

According to Leffa [6], the term authoring tool (AT) refers to a type of software which objectives to generate Learning Objects such as texts, images, videos, audios, among others. These tools are both offline or online and have as targeted audience students and teachers. The author states that the preference for these systems is due to their ability to offer a quick and easy way to create quality educational content, requiring only the input of knowledge and creativity.

Among the advantages of ATs should be cited the high level of interaction the user is likely to have with the object – making learning easy –, and the low cost of material generation [6]. However, it is important to note that despite these advantages, the produced objects should not replace the teacher, but rather complement the work in the classroom. There are several ATs: ALOHA, Ardora, CourseLab, PALOMA and eXe-Learning, for instance.

It is noteworthy, though, that most of those tools have hardware and/or software compatibility issues, require installation of additional applications, do not use the OBAA standard or are not free or user-friendly.

3 Proposed Tool

This document proposes an authoring tool for learning objects capable of generating educational content and its metadata in the OBAA format, and performing the packaging with the SCORM standard, as well as saving the created object in the tool's own repository, conferring user's wish. Its main characteristic is to differentiate itself from other applications for not requiring any kind of installation and being free, intuitive and easy to use. For this, the system takes advantage of the Web platform, being a prerequisite for its use just any web browser with Internet access.

3.1 Project

The tool here presented is intended to be a mixture of repository with authoring tool, presenting itself simple, easy to use and intuitive. For this, it has been pursued to take advantage of a clean and friendly interface as well as presenting the possibility to change language and the use of a feature which automatically adjusts the layout in accordance to the screen resolution. Consequently, the user should not spend too much time looking for where to click or trying to understand what must be done. Furthermore, by changing the layout, functions are released or hidden, avoiding, for example, that someone using the system in a mobile phone access, by mistake, the option of content creation, which requires text or image input, formatting... in short, a task that would rarely be accomplished through that type of device.

Of the repository and authoring tool functions, the system comprises four independent components, which can be used singly or in combination, as the user wishes. Such divisions are detailed below.

- *Educational material making:* the tool features an editor capable of performing content creation based on text and media (audio/video/image), the output of which is the metadata filling screen. The editor can be accessed through the main page on desktop computers and laptops.
- *OBAA metadata filling:* following the use of the editor or directly through the home screen, for desktop computers and laptops, it is presented a metadata form according to the OBAA standard. The form handling performs the creation of the pertinent files and the packing in SCORM, and stores the object in the tool's database. After submitting the form, the user will receive a ZIP file (the very Learning Object itself) via direct download and will be prompted whether to keep or not the object in the repository.
- *SCORM packaging:* packaging occurs automatically after submitting the metadata form or, if the LO has been previously made, it can be accessed through the home screen on desktop computers, laptops or tablets. If the latter, it will be shown a screen where the user uploads his LO and the system converts it to SCORM

format. In both cases, the LO will be saved in the repository and the user will be prompted if they want to keep it that way or remove it, in addition to receiving their packaged LO via direct download.

- *Search:* to any device used, one can perform a search in the repository database from the main screen. For smartphones it is only possible to execute textual search, while for tablets, laptops and desktop computers it is possible to make use of a sliding menu of images, representing different themes-filters for search. The output is a screen showing the search results, with brief summaries of found LOs, link to details of each LO and filters to refine results. The details page is also ruled by the resolution of the device in use: mobile is only given the option to view the LO, while for the other devices one can download the LO without SCORM, with SCORM or simply view it.

The prototype of screens and their relationships, as well as the overall class diagram, were proposed by the project's mentor teacher, and then adapted as appropriate during the development process.

3.2 Development

The development process began with the study of needed languages and technologies: PHP, HTML, CSS, JavaScript and Database/MySQL; as well as Object Oriented Programming, Model-View-Controller and OBAA and SCORM standards. Next, from the model proposed by the mentor teacher, were developed the tool prototype layouts for the different devices and basic versions of the most significant pages, including: the form for the metadata descripting, the editor for creating the educational content and the home page with the sliding menu used to filter search and the textual search field. Consequent to it happened the database modeling and generating, followed by the construction of the PHP classes which would be responsible for the objects representing the metadata and for the translation function. It is emphasized here that all mentioned components suffered, constantly, changes to meet all specifications and unexpected errors.

It has also been worked on the XML file, done through PHP's DOMDocument class, which allows a XML file in accordance to the OBAA model. It is utilized the ZipArchive class as well, to compress the XML file along with the educational midia and the HTML page which references said midia.

The SCORM components (and several other minor details) were not, yet, worked on as the tool is just in early stage of development.

3.2 The Tool Working

In the tool's current version, considered by the authors as early stage of development, are presented, fully operative, the following features (already termed in previous sections): metadata form with XML file generation, creation of educational content through a basic editor, storage in the database, search from textual terms and translation and automatic layout change functionalities.

Figures 1 and 2 illustrate the operation of the tool here shown. Figure 1 shows the main screen with the search system and three buttons: one for each function of the tool. In Figure 2 it can be seen the form for metadata filling as the OBAA standard.

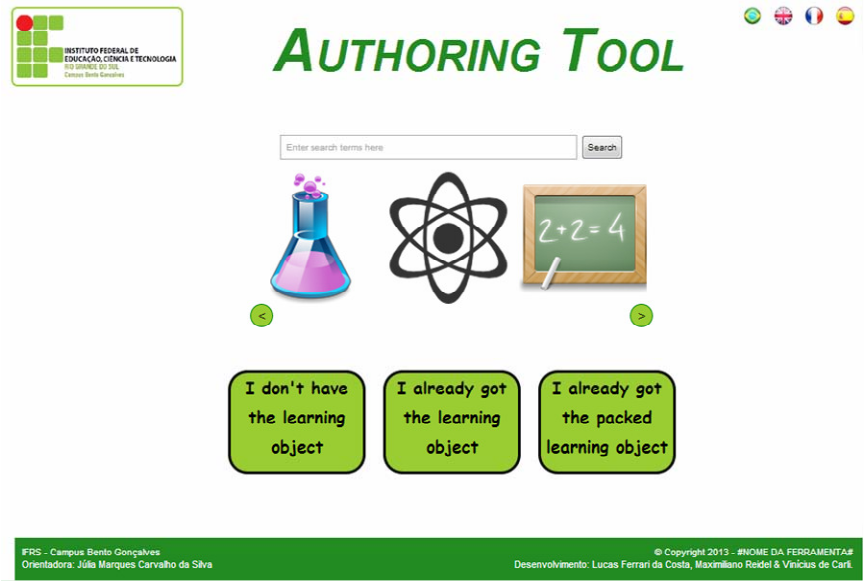


Fig. 1. System's home screen.

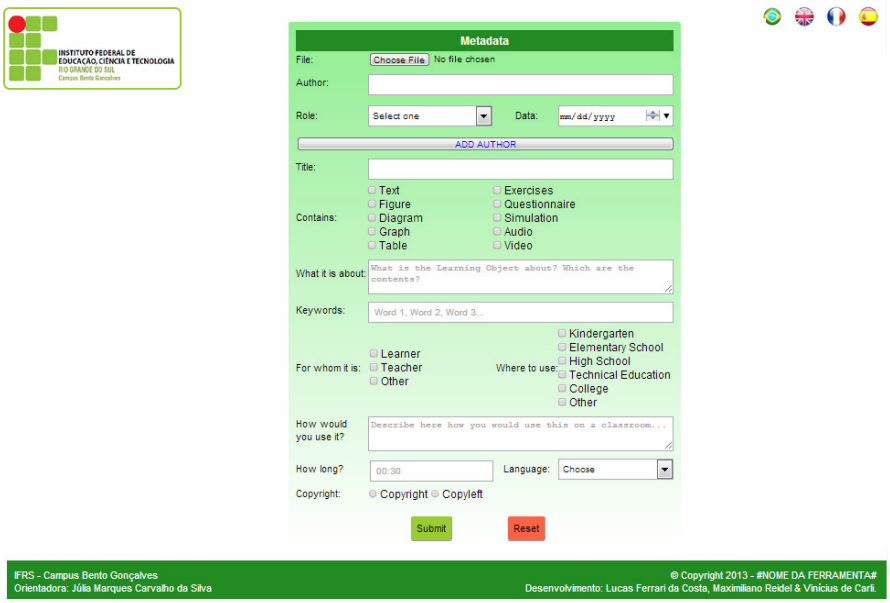


Fig. 2. Screen with the form for OBAA metadata descripting.

4 Conclusion

Though little known, the learning objects (LOs) have become popular and necessary, perpetrating imperative the presence of tools to assist in their creation, maintenance and dissemination. Even more important is the existence of applications capable of generating objects in accordance with the Brazilian standard, in order to encourage and disseminate its culture.

With the main features - such as searching for objects from the repository, creating LOs with the editor and inserting objects in the database with the metadata cataloging - already operating, it is tangible that the proposed system, even in early stage of development, meets the accentuated profile.

Next, it is expected to improve the search function with filters and perform the LO conversion to SCORM. It is planned to, in the future, improve the LOs editor and the system layout.

References

1. Wiley, D. A.: Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. Digital book, http://reusability.org/read/#1 (2000).

2. Silva, J. M. C.: Análise Técnica e Pedagógica de Metadados para Objetos de Aprendizagem. PhD thesis, Universidade Federal do Rio Grande do Sul (2011).
3. Vicari, R. M., Bez, M., Silva, J. M. C., Ribeiro, A., Gluz, J. C., Passerino, L., Santos, E., Primo, T., Rossi, L., Bordignon, A., Behar, P., Filho, R., Roesler, V.: Proposta Brasileira de Metadados para Objetos de Aprendizagem Baseados em Agentes (OBAA). Technical report, Universidade Federal do Rio Grande do Sul (2010).
4. Advanced Distributed Learning, http://adlnet.org/scorm.
5. Tarouco, L. M. R., Dutra, R.: Padrões e Interoperabilidade. In: Prata, C. L., Nascimento, A. C. A. A.: Objetos de Aprendizagem: Uma Proposta de Recurso Pedagógico. MEC/SEED 2007, pp. 81-92. Brasília (2007).
6. Leffa, V. J.: Uma ferramenta de autoria para o professor: o que é e o que faz. In: Letras de Hoje. PUCRS 2006, pp. 189-214. Porto Alegre (2006).

Incorporating Virtual Activities in Higher Education: A Mathematical Model for Describing Teachers According to their Skills

Lucia Rosario Malbernat

Department of Systems at the CAECE Mar del Plata University, Argentina
lmalbernat@ucaecemdp.edu.ar; lmalbernat@gmail.com

Abstract. Teachers must innovate in their practices to incorporate virtual activities at the university. They must develop teaching skills related to their own preparation and attitude towards virtual education.
This paper presents a model designed to quantify some manifestations of the preparation and attitude which are necessary to create environments for online distance education. This model has been applied to data processing from CAECE and UNMDP Universities. Some conclusions are presented here.
The following indicators were taken to define faculty preparation: level of use of ICT, training and experience on virtual education and mastery of computing tools. In order to calculate the attitude towards virtualization it was necessary to define the following indicators: level of interest in the use of ICT, interest in virtual training, stance on relationship with ICT and stance on virtual education.

Keywords: distance education, university innovation, virtualization, ICT, teaching skills, indicators

1 Introduction

Incorporating ICT in higher education is a case of innovation that could not exist without the development of backgrounds and environments for technology-mediated education.

These innovation processes mediate education so that, between an student and the content to be learned, instead of having teachers who transmit information, there may be a learning facilitator to guide the student in their search and technological means to provide him not just a lot of information, but also diversity of motivations and forms of communication.

Technology-mediated learning can free the teaching process from temporal-spatial constraints and encourage academic events in which it is feasible to mediate spatial and temporal distance between teacher and student by interactions. Despite this, they continue to establish social bonds.

Therefore, when information technologies are used in education, in addition to students and teachers- technological media is involved. Teachers also fulfill non-traditional roles as they may participate asynchronously with the student. In spite of that, they maintain strong interactions. In this context, the feasibility of incorporating virtual activities into undergraduate courses considering teaching skills has been and is being studied.

This paper presents a model proposed to transform direct scores, obtained from a questionnaire designed ad hoc, into indicators. These new scores provided information to quantify the two standard measures used to characterize teachers: Preparation and Attitude. The teacher questionnaire was adapted from data collection instrument described in [1].

A case study developed at CAECE University (UCAECE), Systems Department, located in Mar del Plata, Argentina is presented for the purpose of sharing the results of processing the collected data but they are currently being contrasted with similar data taken at the Mar del Plata National University (UNMDP), Faculty Economics and social Sciences.

In addition, information obtained by applying the proposed model has been taken as input to the segmentation algorithm described in [2]. The segmentation algorithm was designed with the aim of bringing together teachers according to their preparation and innovative approach for incorporating virtual activities. The proposed segments are: Innovators, Indifferents and Resisters. Uncertainty in decision-making related to selection of teachers, incorporation of online activities into courses and teacher training (such as described in [3]) could be reduced with the analysis of said information.

1.1 Incorporating Virtual Activities in Higher Education

Casas Armengol [4] believes that innovation and virtualization of universities are essential instruments to boost great scientific social changes to effectively progress towards the future knowledge society. Also, according to Garcia et al. [5] distance education has been, since its beginning, the modality that has shown greater readiness to take technological innovations.

Thus, very solid proposals for facilitating change processes and defining factors and approaches designed to achieve widespread use of technology (in relation to the definition of roles, functions and track record) are very abundant.

Writings, studies and research related to teaching skills [6], [7], [8], [9], [10], [11] are frequently made. In 2008, UNESCO published the ICT competency standards for teachers to provide guidance for planning teacher education programs and selecting courses to prepare them for the student's technological training [12].

Moreover, the attitude concept has traditionally been defined as ~~a~~ willingness to respond either favorably or unfavorably towards an object, situation or event [10]. Training and instruction can help improve this attitude. It is understood that the knowledge necessary to incorporate ICT covers various aspects. These aspects are some teaching skills which define the teacher's preparation (training, experience, expertise, etc.) and attitude (intrinsic and extrinsic interests, opinion, etc.) to perform online activities.

1.2 Quantifying Indicators of Teacher's Preparation and Attitude

In psychology, education and social sciences there are aspects which are measured but are not physical or directly observable [13].

The measurement of an attribute by a test gives a score-direct but a person's raw score on a test is not directly interpretable if it is not compared, for example, with the performance of people from the same group [14]. It has no meaning in itself. It becomes meaningful when it is compared with standard tables and previously constructed scales with scores obtained from the group by applying the test [13]. Thus, a subject's score on one aspect (indicator) may be compared (in certain scale) with people who make up the group scores [14].

In the research reported in this paper, in addition to analyzing the direct scores obtained using the questionnaire described in [1], we propose a model for transforming direct scores into derived scores, which normalizes the two measures taken to characterize teachers (see Table 1). The model which allowed to calculate and to assign a value representative of the dimensions Preparation (P) and Attitude (Q) of each teacher by calculating their respective indicators is described below.

Table 1. Indicators to calculate P and Q composite indexes.

Preparation **(P)**	Attitude **(Q)**
(R) Level of ICT use	(U) Interest in the use of ICT
(O) Mastery of tools	(I) Interest in virtual training
(F) Training in Virtual Education	(N) Stance on relationship with ICT
(E) Experience on Virtual Education	(G) Stance on virtual education

2 Model to Quantify Teacher Preparation and Attitude

Let P be Preparation index and let Q be index Attitude to incorporate online activities in teaching, δ the set of teachers who are studied, π the set of quantitative p_i indicators, as understood in this paper, teacher Preparation (P), and θ index, the equivalent for Attitude, with $q_i \in \theta$, may be defined by extension as π = {R, O, F, E} and θ = {U, I, N, G} respectively.

The following describes proposed calculus to obtain a representative value for each teacher, considering all their quantitative indicators, based on data collected through the survey of opinion. With their application you get the ζ set of ordered pairs of the form (P, Q) representing an element of the set of teachers.

2.1 Composite index from Teacher Preparation (P):

The teacher's preparation is defined by P index and it is calculated with Equation 1. It can reach the maximum 10 decimal points. Each indicator can bring his a maximum score of 2.5 points:

$$P = R + O + F + E\ . \quad (1)$$

With P, Preparation index, R, Levels of ICT use; O, Mastery of tools, F, Training on Virtual Education, E, Experience with virtual education and 0 >= P <= 10; 0 >= R <= 2,5; 0 >= O <= 2,5; 0 >= F <= 2,5; 0 >= E <= 2,5.

ICT use level (R). Classifications used to quantify R and U indicators were taken from the CBAM (Concerns-Based Adoption Model). IT is described in [4].

CBAM includes seven levels teachers go through during the process of incorporation of technology. Table 2 shows these levels, and an additional one to include teachers who don't use ICT.

Table 2. Levels of ICT use (R)

Levels of Use	Behavioral Indicators of Level
7, Renewal	Teacher seeks to improve the use of ICT. He/she reevaluates his/her use and examines new innovations as better options
6, Integration	Teacher is making deliberate efforts to coordinate with colleagues in using innovation to improve results
5, Refinement	Teacher is considering implementing changes in the use of ICT to improve learning outcomes of his students
4, Routine	Teacher performs a basic use of ICT because he has an established pattern of use; changes are specific
3, Mechanical	Teacher has focus on immediate and mechanical aspect of ICT, he/she uses it repeatedly and at their own convenience
2, Preparation	Teacher is prepared to use ICT
1, Orientation	Teacher is learning what are TIC about, he/she begins to discover ICT
0, Non-Use	Teacher is taking no action; he/she does not do any activity with ICT

This classification is useful for monitoring the level educators are going through in relation to interest in the use of ICT and the degree they use it effectively ~~it~~ [15].

Teachers go through the levels sequentially. Therefore, the maximum score (R = 2.5) of the indicator is linked to the most comprehensive selection (Ri = 7) and corresponds to the maximum level attained. Level 0 provides no score (Ri = 0). Thus, the chosen r_j can take an integer value in the range [0, 7], which coincides with the highest level achieved by the teacher.

$$R = \frac{2{,}5 * r_j}{7}. \quad (2)$$

Whit 0 <= r_j <= 7.

Mastery of tools (O): The teacher may indicate his/her proficiency in the use of each tool to be very appropriate, appropriate, regular, inappropriate, very

inappropriate and may indicate "do not know or no answer". It is understood, therefore, that the maximum contribution that each item can make to teacher preparation is verified when the “very appropriate” option selected for a specific tool.

The minimum contribution ($O_i = 0$) corresponds to the mastery of tools “very inappropriate” choices (or “do not know” / no answer).Therefore, intermediate options which refer to appropriate mastery, regulate and inappropriate provide intermediate values for preparation weighted as 0.75, 0.5 and 0.25 respectively. Consequently, the estimate of this indicator can be defined as normalization of the sum of the values v_j of each items which contributes to the indicator O, weighted:

$$O = \frac{2{,}5 * \sum_{j=1}^{11} (v_j * t)}{11} \quad (3)$$

with $v_j = 0$, if the chosen alternative in the item is "Do not know / no answer", or 1 in any other election, and t the weighting factor for the election as following: Very appropriate, 1; Appropriate, 0.75, Regular, 0.50, Inappropriate, 0.25; Very inappropriate, 0.

Items on which the teacher should define his/her mastery of tools are: browsing institutional virtual campus; reporting news, files or sites in the institutional virtual campus; obtaining information and resources via the Internet; using e-mail for sending and receiving messages; using e-mail for sending and receiving enclosed files (attachments); creating groups or rules; being involved in discussion milieu, opinion forums and blogs; being involved in chat rooms; administrating and managing blogs; creating office documents; creating of multimedia documents.

Training on virtual education (F). Since each choice sets up a contribution, it adds one value for each chosen subject (v_i). That is, the maximum score that f_i can bring to the teacher´s preparation –previously defined– (F = 2.5) corresponds to the 6 values of $v_i = 1$, which is the case in which the teacher has been trained in the 6 issues referred into the polls, while the lowest score ($R_i = 0$) corresponds to the teachers who have not been trained in any of them.

$$F = \frac{2{,}5 * \sum_{j=1}^{6} v_j}{6} \cdot \quad (4)$$

with $v_j = 1$ if the option was chosen by the teacher and $v_j = 0$ otherwise.

The options the teacher can choose for this indicator and for interest in training (I) are the following: use of ICT and/or media; methodologies that can improve teaching practice if using ICT; techniques of learning facilitation for virtual education; alternative assessment methods appropriate for when using ICT and instructional design, management and/or planning of virtual education.

Experience on Virtual Education (E). The v_j items also provide one value for every positive teacher e_j choice. Therefore, if the teacher did not choose any option, e_j value is 0 and if the teacher chose all options, it will be E = 2.5. The latter is the case where the 7 v_j values are equal to 1.

$$E = \frac{2{,}5 * \sum_{j=1}^{7} v_j}{7} \cdot \tag{5}$$

With $v_j = 1$ if the option was chosen by the teacher and $v_j = 0$ otherwise.

The options the teacher can choose for this item are: He/she has attended courses on virtual education (distance learning, online, open, e-learning); has attended training courses not related to virtual education but virtually dictated; has taught courses on virtual education; has taught courses unrelated to virtual education but performed virtually; has performed as a learning facilitator in online courses; has designed or planned courses delivered virtually or has managed them in some way; has actively participated in virtual congresses (with at least 20 hours of virtual activities).

2.2 Composite index from Teacher Attitude (Q):

The Q composite index is calculated with Equation 6. It can reach the maximum 10 points in the decimal scale. Each indicator can bring his a maximum score of 2.5 points.

$$Q = U + I + N + G \, . \tag{6}$$

With Q, Attitude index; U, Interest in the use of ICT; I, Interest in virtual training; N, Stance on relationship with ICT and G, Stance on virtual education and 0 >= Q <= 10; 0 >= U <= 2,5; 0 >= I <= 2,5; 0 >= N <= 2,5; 0 >= G <= 2,5.

Interest in the use of ICT (U). The maximum score (U = 2.5) of the U indicator corresponds to u_j = 7; it describes teachers who reached the highest level. This indicator has a similar treatment of the scale "Levels of ICT use". The lower contribution (U_i = 0) corresponds to the choice made by those teachers who neither know nor have any interest in ICT.

$$U = \frac{2{,}5 * u_j}{7}. \quad (7)$$

Whit 0 <= u_j <= 7 that coincides with the level reached by the teacher.

Table 3 shows levels of interest in the use of ICT and their descriptions.

Table 3. Levels of Interest in the use of ICT

Levels of Interested	Behavioral Indicators of Level
7, Refocusing	The teacher has ideas about how to improve the use of ICT and how he/she can do a better implementation of them
6, Collaboration	Teacher discusses how to collaborate with colleagues involved with ICT
5, Consequence	Teacher begins to consider the impact that ICT can have on student learning
4, Management	Teacher has concerns about the administrative and logistical challenges posed by ICT; they consume his/her time
3, Personal	Teacher asks himself what impact ICT can have on his/her person in relation to time and his/her own abilities
2, Informational	Teacher, at this level, wants to know more about ICT
1, Awareness	Teacher knows about ICT but they don't generate him/her any concern
0, Without Awareness	Teacher has not yet begun the process of innovation

Interest in virtual training (I). This indicator assigns a score to current or past interest in training. Each v_j positive choice for the 6 statements of the questionnaire, provides a point i_i. Each question has been asked in a "mirrored" way with 6 items consulting on current training (v'_j), corresponding to the indicator O.

The negative choices of the I indicator were considered positive when the equivalent in training already performed was positive.

Consequently, the indicator I carries a value of 2.5 points when the 6 values v_j (or its equivalent v'_j defined by questionnaire items by the indicator O) are equal to 1 because they have been selected by the teacher (see Equation 8)

$$I = \frac{2{,}5 * \left(\left(\sum_{j=1}^{6} (v_j \quad OR \quad v'_j) \right) \right)}{6}. \quad (8)$$

with $v_j = 1$, $v'_j = 1$ when the option was chosen by the teacher, and $v_j = 0$, $v'_j = 0$ otherwise and OR logical operator truth table V_j OR V'_j.

Stance on relationship with ICT (N): It is understood that the contribution of each v_j item to the Q index is defined by the option the teacher chose for each one.

The "Total agree" choice adds 1 point, the intermediate options (agree, neither agree nor disagree and disagree) contribute to N index 0.75, 0.5 and 0.25 points respectively, while "Total disagreement" brings 0.

Therefore, the calculation of N will be defined as the normalization of the sum of the V_j weighted. See equation 9.

$$N = \frac{2{,}5 * \sum_{j=1}^{6} (v_j * t)}{6} \cdot \tag{9}$$

with $v_j = 0$, if the selected option in the item is "Do not know / no answer", or $v_j = 1$ in any other selection and t, the weighting factor of the election, according to the same weight as the R indicator.

The teacher responds to the following items: If his/her own computer knowledge is appropriate for his/her needs and if it is suitable for the use he/she wants to give it; if his/her ICT skills meets current personal expectations; if his/her attitudes towards the use of ICT is positive; if teachers can obtain benefit from virtual education because they can better manage their time; if it may be beneficial for the teacher to dictate blended courses; if dictating virtual courses can bring some benefit or utility (professional development, work from home, etc.) for the teacher.

Stance on virtual education (G). G indicator has a similar treatment N indicator. The maximum score of G indicator is given to the choice made by the teacher on the maximum degree of agreement (total agreement) for v_j items associated to the indicator. Intermediate options are valued with the respective weighting factors 0.75, 0.5, 0.25 and 0.

$$G = \frac{2{,}5 * \sum_{i=1}^{7} (v_j * t)}{5} \cdot \tag{10}$$

with $v_j = 0$, if the selected option in the item is "Do not know / no answer" or $v_j = 1$ in any other election and t, the weighting factor of the election.

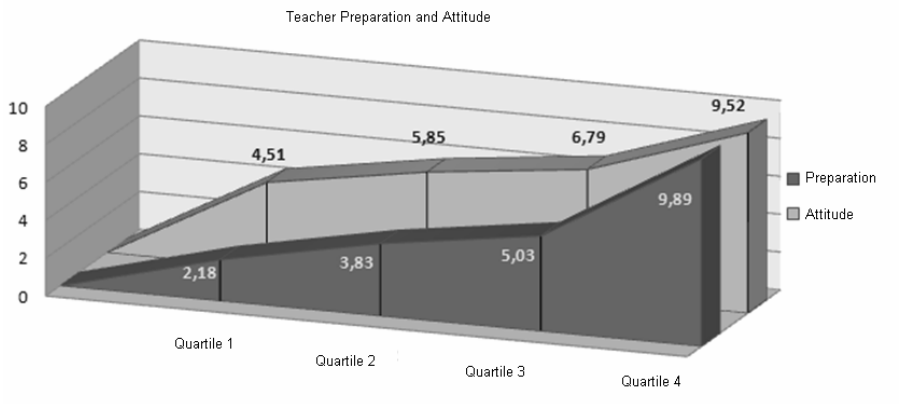
The items on which the teacher has to deliver an opinion are:

the quality of virtual education can be equivalent to classroom education using appropriate methodologies; the teacher can achieve academic interventions in virtual education such that they are equivalent to interventions performed in classroom education; Educational model based on the use of ICT can help the professional development of teachers; Educational model based on the use of ICT can help to the teaching practice in the classroom; Distance learning can provide benefits to teachers in relation to time management.

3 Conclusions

The model presented in this paper has allowed arriving to relevant information for decision making related to teachers, so as to promote appropriate environments for distance education. Its aim is to quantify certain manifestations of preparation and attitude to incorporate virtual activities in higher education.
The following information emerges from data analysis on which the model is applied. It is currently being compared with data obtained in other study house. For further information see [3].

The graph below describes preparation and attitudes of teachers, separated by quartiles. The second quartile (50th percentile) coincides with the median value. Denotes a low preparation (measured in 3.83, when the maximum possible value of P is 10) and a positive attitude (5,83).



Graph 1 – Preparation and teaching attitude for incorporating virtual activities

Both the teacher preparation average and the teacher attitude average are slightly higher than the averages reported by auxiliary teachers survey responses.

It is notoriously high the attitude of students playing the role of assistants with

respect to most of the sample, in most disciplines. The following table presents summary information for the statistical analysis carried out for indicators which describe the P index.

Table 4. Summary of statistical analysis of P (Preparation)

Statistical Analysis	Level of ICT use (R)	Mastery of tools (O)	Training in Virtual Education (F)	Experience on Virtual Education (E)	**Prepara-tion (P)**
Mean	1,25	1,52	0,59	0,44	**3,80**
Median	1,43	1,59	0,42	0,36	**3,83**
Mode	0,00	1,82	0,00	0,00	**4,32**
Deviation	0,92	0,55	0,71	0,48	**2,00**
Maximum	2,50	2,50	2,50	2,50	**9,89**

Statistical Analysis	Interest in the use of ICT (U)	Interest in virtual training (I)	Stance on relationship with ICT (N)	Stance on virtual education (G)	**Attitude (Q)**
Mean	1,61	1,21	1,49	1,20	**5,51**
Median	1,79	1,07	1,56	1,31	**5,85**
Mode	1,79	0,71	1,88	1,50	**6,46**
Deviation	0,64	0,66	0,61	0,64	**2,06**
Maximum	2,50	2,50	2,50	2,50	**9,52**

Note that the range of valid values of the indicators make possible calculation of P and Q is [0, 2.5] and the range of valid values of P and Q index is [0, 10].

Finally, as a corollary, it is stressed that P & Q indexes were taken as segmentation variables for classifying each teacher [2]. Three clusters have been defined (Innovators, Indifferents and Resistors) because it was understood teachers at least can be classified into three categories [15]: Teachers with positive attitude towards ICT who improve the standards for teaching and learning, teachers who assume neutral position regarding ICT use in education and teachers with explicit negative attitudes toward new technologies.

At CAECE University, it emerges from the application of the algorithm that 17.39% of the overall sample was included in the cluster of Innovators, the wide majority of 53.62% was located in the segment of Indifferents and 28.99% of them fell in Resisters group-teachers with negative attitudes towards new technologies. Similar tendencies were found at the University of Mar del Plata applying the algorithm. This information was used for decision making related to teacher training and planning to open virtual classrooms providing information to design a training plan and to anticipate the amount of virtual classrooms requested. This will be shared in future presentations.

References

1. Malbernat, L.R.: TICs en educación: competencias docentes para la innovación en pos de un nuevo estudiante. VI Te&ET. (2011)
2. Malbernat, L.R.: Incorporar actividades virtuales en educación superior: algoritmo de segmentación de docentes según sus competencias. XV WICC (2013)
3. Malbernat L.R. Innovación en educación universitaria: Factibilidad de incorporar actividades virtuales según las competencias docentes M.S. Thesis, Mar del Plata National University, Argentina (2012).
4. Casas Armengol, Miguel. Nueva universidad ante la sociedad del conocimiento. Revista de Universidad y Sociedad del Conocimiento vol. 2, Nº 2 UOC (2005)
5. García Aretio, L. Ruiz Corbella, M. & Domínguez Figaredo, D. De la educación a distancia a la educación virtual. Barcelona, España: Ed. Ariel. P 42 (2007)
6. Beneitone, P., Esquetini, C. González, J. Maletá, M., Siufi, G., & Wagenaar, R. Reflexiones y perspectivas de la Educación Superior en América Latina. Informe Final – Proyecto Tuning- América Latina 2004-2007. Bilbao, España. (2007)
7. Zabalza M. Competencias docentes del profesorado universitario. Calidad y desarrollo profesional, (2a ed.) Madrid, España: Narcea S.A. de ediciones (2007).
8. Gutierrez Porlán I. Competencias del Profesorado Universitario en relación al uso de Tecnologías de la información y la comunicación Análisis de la situación en España y propuesta de un modelo de formación. Tesis doctoral. Tarragona. España. Pag 447 (2011)
9. Prendes Espinosa, M. P. Competencias TIC para la Docencia en la Universidad Pública Española: Indicadores y Propuestas para la definición de buenas prácticas: Programa de Estudio y análisis. Informe del Proyecto EA2009-0133 (2010)
10. Álvarez, S., Cuéllar, C. López, B., Adrada, C., Anguiano, R., Bueno, A. et al. Actitudes de los profesores ante la integración de las TIC en la práctica docente. Estudio de grupo de la Universidad De Valladolid. Edutec-e. No. 35 / marzo. ISSN: 1135-9250. P. 3 (2011)
11. Prendes Espinosa, M. P, & Castañeda Quintero, L. Universidades Latinoamericanas ante el reto de las TIC: Demandas de Alfabetización tecnológica para la docencia. Comunicación proyecto A/018302/08, Estudio de las competencias y demandas formativas en TIC de los docentes de las Universidades bolivianas y dominicanas. Universidad de Murcia. Murcia: España (2010)
12. UNESCO. Estándares de Competencia en TIC para Docentes. París, Francia: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2008)
13. Aliaga Tovar, J. Psicometría: Tests Psicométricos, Confiabilidad y Validez. Capítulo 5 del libro Quintana, A. Montgomery, W. (Eds.) Psicología: Tópicos de actualidad. Liman Perú: UNMSM. P. 86-88 (2006)
14. Abad, F., Garrido, J. Olea, J. & Ponsoda, V. Introducción a la Psicometría. Teoría clásica de los test y teoría de la respuesta al Item. España: UAM. P 119 (2006)
15. UNESCO Las tecnologías de la información y la comunicación en la formación docente. Guía de planificación. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, p 169 (2004).

Interoperability in virtual world

Leandro Rosniak Tibola[1,2], Liane Margarida Rockenbach Tarouco[2]

[1] Regional University Integrated Alto Uruguai Missões, Assis Brasil Av. 709, 98400-000, Frederico Westphalen, RS, Brazil
[2]Graduate Program Informatics on Education, Federal University of Rio Grande Sul, Paulo Gama Av. 110, 90040-060 - Porto Alegre, RS, Brazil
tibola@uri.edu.br, liane@penta.ufrgs.br

Abstract. Recently we have seen a growing use of virtual worlds (VW) in many areas like marketing, e-commerce, games, social interaction and education. The virtual worlds offer many resources to engage their users (named avatar) like freedom of movements, teleport yourself to other local, communicate with other inhabitants (both text and voice messages), capacity to create, modify and destroy objects and the possibility of programming behaviors to these objects via scripts. The amount of resources "in" the world is great. This measure will be bigger if adding the resources "out" the world. Both resources (in and out) need ways to communication. Our interest on VW applications to education goals, rises the necessity of understand the model, protocols and ways of communication these "virtual worlds" with the "real worlds".

Keywords: virtual world's communication, Open Sim, Second Life, real world to virtual worlds, R2V.

1 Introduction

If we can say: the life imitates video, and then we can say that the real world going to virtual world and vice versa. A long time the movies carry us to worlds where the reality is overtake, no are limits to interactions with objects and inhabitants and the human communicate with computer from some way.

Experiences that involve the human brain connection with computers and virtual reality are the recurring theme in the movies. Many current research to connect real and virtual worlds, already presented in the cinema, like communication (as showed in the 1992 movie The Lawnmower Man), teleportation (as depicted in the 1982 movie Tron), one world affecting the other (as viewed in the 1999 movie The Matrix) and the human feelings and behaviors (as showed in the 2009 movie Avatar).

Recently, games are using the real devices to improve virtual interaction player-environment-others players. Two well-known samples are Nitendo Wii and Microsoft Kinect. Wii use sensor-based technology to scan accelerometers and movement sensors to synchronize the motion of the game character with the real motion of the gamer [1]. Kinect apply camera and real-time vision processing to recognize user

movements and microphones that calibrate the player voice to control an avatar or an object in the virtual world according to the movement of gamer [2].

Besides, VW technology has been applied in many areas with differing goals. The companies using the lands to publicize and more deep sense, test the acceptance their products or services [3]. The lands are fertile ground to create, innovate and disclose objects with the brand, concepts and values of the companies, reinforcing their influence with the costumers. Others actions can involve sales, discounts, awards and competitions. The flexibility and velocity to update the objects and lands with the costumer's desire are great reasons to firms be in VW.

The individuals, through her/his avatar, can market in these worlds too [4]. At the same time that real market happens out of virtual worlds, the virtual commerce happens in virtual worlds. People create and sale objects, terrains, avatars and all items there are in these environment. People might want just fun, meeting other people and play in the world without worrying with theoretical or technical issues [5]. A great number of people have these behaviors in the VW's. Make friends, know different cultures, be accepted in the groups, share interests and be part of something are causes to persons leave aside technical aspects and focus in personal relationship.

The education is an area that receives a lot of effort to use the virtual worlds resources. Attractive visual, flexibility, availability, simplicity, interaction, collaboration, cooperation, media support, freedom creative, and free virtual worlds servers are some arguments to teachers employ VW in their classrooms and students want access it. According [6], VWs are good resources to achieve educational goals because they have persistence (effects of actions remain in the environment even if the user leaves), access and availability 24 hours a day, seven days a week; allow social interactions; their 3D graphical environments would improve interaction and sense of realism; is possible to see, hear and touch virtual objects as well as create, edit and manipulate them as if they were physical objects. These technologies also allow teachers and students the use of innovative learning strategies: practical training, group work, discussions, field practices, simulations, and visualizations of concepts.

Many researches about communication among real and virtual worlds and ways to exchange information between these environments are in progress. In [7] is proposed a method to transform sensed information from real world to standardized XML instances and to control virtual world objects. Sensor data imported into the multi-user 3D environment that mirror a real-world chocolate factory and its processes are presented by [13]. The problems and difficulties of data exchange between virtual worlds and the real world, comparison between different fundamental approaches of realizing communication channels and outline a general method for designing interfaces between virtual worlds and real-world data spaces are described by [8]. Provide an understanding of virtual wealth, its related virtual world, and its relations to real world; arguing that the realization of virtual wealth in both virtual and real worlds is necessary and possible is aims of [9].

The need for ontology services, the several approaches for associating ontology concepts with objects and locations, discussion how to populate common-sense ontologies using data harvested from real and virtual worlds are showed in [10]. The virtual campus project which builds a 3D virtual campus of the Chinese University of Hong Kong (CUHK) in networked virtual worlds and where users can immerse in

virtual CUHK with an avatar and experience virtual education and other diverse campus activities is reported in [11]. The collect of information about network traffic to balance the client-server response in Massively Multiplayer Online Games (MMOGs) is depicted in [12].

Understand the influence of real world house price on virtual world land price with data collected from SecondLife.com and federal housing finance agency of USA and finding that land price in virtual world is significantly associated with real world land price is account by [14].

Two interesting proposals involving haptic stimulation could be viewed in [15], where one of the virtual avatars kisses the other in Second Life, an event is triggered and this event is decoded by our system to send haptic based kiss to the real user via the Bluetooth-enabled neck piece hardware. The other proposal [18], aims facilitate the communications of emotional feedbacks such as human touch, encouraging pat and comforting hug to the participating users through real-world haptic stimulation by the development of a prototype that realizes the virtual-real communication through a haptic-jacket system. Both works indicate some of the potential applications like distant lover's communication, remote child caring, and stress recovery.

In [16] is presents a novel system for detecting the real-world activities (events) using visual means (surveillance systems) over an area of the building and sending relevant information to the Virtual World Servers for recreating the events in Virtual World representation of the same area. The human moves like entry, exit, standing, walking, bending are captured and sent to virtual world server via HTTP messages.

Infer the human activity from environmental sound cues and common sense knowledge as an inexpensive alternative to expensive sensors and discuss the challenges to implement such a system from the signal processing and agent based system was writing by [17].

Following the concept of Internet of Things (IoT), some researches employ common devices to collect data from real world and send it to virtual worlds. In [19] is showed a scenario where context data collection from mobile devices can be used for augmenting virtual worlds with real-life data. Life-logging elements are used to control an avatar, in a virtual world, as a way to replay experiences. The mobile smat phone equipped with the context collection daemon which constantly monitoring the location of the device from the Global Positioning System (GPS) and the songs played by the standard media player of the phone. This information is send to Virtual World Server via HTTP web interfaces.

Using smart phones too, [20] presents a proposal for the use of 3D worlds to enhance the interface of mobile-learning applications. Some specific test results are shown for every component available to construct 3D worlds and the result is an expanded interface where more information is displayed in the same space related to the subjective 3D perspective.

Expanding researches possibilities, [21] describe how technologies like Ubiquitous Intelligence, Cyber-Individual, Brain Informatics, and Web Intelligence can be integrated together and fit into a seamless cycle like the one proposed in the Wisdom Web of Things (W2T). To achieve this goal, are showed two cases studies with communication between real and virtual worlds.Thus, is clear that the uses of VW are growing and expanding in many areas. To full use of virtual worlds, is necessary to communicate real and virtual worlds. This paper explores the communication ways

from real world (material devices and resources) to virtual world (avatars, objects and scripts) using the Open Simulator server, also called OpenSim [22]. OpenSim was used because it is very popular and free software. Moreover, it is applied in many researches, has wide documentation, wikis, and blogs, has users and developers' large base and is compatible with Windows, Linux and Mac OSX platforms. We think that our examples should work at Second Life server [23] too, but they have not been tested yet.

The reminder of this paper is organized as follows: in Section 2 are described the strategies of interoperation in a virtual world, Section 3 shows the results achieved and Section 4 ends with conclusion and future works.

2 Interoperation in a Virtual World

This section lists the communication protocols available in OpenSim and describe applications of some theirs. As are many possibilities, we stay with those used in subject cited below.

2.1 Communication in the World

The user can communicate with others users across her/his avatar or by script programming inserted in the object. Most simple way is the chat. Chat is available at bottom part of the viewers and allow to sending broadcast message typing the text into chat box. This communication utilizes the channel 0 to chat to all nearby avatars and objects. Text sent by chat through channel 0 can only be perceived by avatars within 20 meters (in virtual world metric) of the sender [25].

Users also can employ the Communication viewer's menu or similar. Most viewers have menu that allow set the chat status (away, auto respond, unavailable), classify chats by friends, contacts, groups and connect he/she thus. The classified people can be found via map's search. Also, lists of gestures that trigger the avatar to animate, play sounds, and/or emit text chat. Voice resource can be turn on/off to detect nearby.

In the virtual world, the avatar communicates with objects by "touch". Touch happens when avatar click on the object and it have a script that execute an action. To [25] touch_start ( ) event is triggered by the start of agent clicking on task. When the avatar touch an object, a stars stream follow from avatar to object, wrapping this object and after that the programmed action (script) is executed.

The objects also communicate with avatars. According [25] the event state_entry( ) enable executing actions when it is detected the proximity of the avatar. Function llSensor( ) is a sensor that detects all avatars/objects and agents with a given name within 15m of the sensor. As soon the object identify who touch it. This information could be check with a list of authorized avatars to realize a specific action.

In addition to avatars, objects can communicate with other objects. One way is the event listen( ). As [26] the listen( ) event handler is invoked whenever a chat message matching the constraints passed in the llListen function is heard (received). The channel the chat was picked up on, the name and id of the speaker and the message are passed as parameter in llListen function. Therefore, an object with function

llListen( ) in event state_entry( ) will 'listen' all messages sent to that channel and execute the corresponding script providing a response .

2.2 Communication Protocols

According [24], OpenSim communication protocols are divided in the following types:

a - Client-Server protocols: these are communication protocols between OpenSimulator and a client or viewer. This is mainly between the viewer and the simulator, though some traffic also flows directly between the viewer and a grid service. The primary protocol here is the Linden Lab viewer protocol, which is carried into UDP messages (in the case of object updates, avatar position updates, etc.) and HTTP based messages via capabilities.

b - Grid service protocols: on a standalone OpenSimulator installation, all communications occurs within process. However, with a grid installation the simulators need to communicate with backend services (asset, inventory, etc.) and this is done over HTTP.

c - Simulator-Simulator protocols: there are some situations in which simulators need to communicate directly with one another: (a) teleports and region crossing (this communication is carried out over HTTP) and (b) instant messaging protocol between users on different simulators (this also covers item giving since this is communicating using the IM infrastructure).

d - Simulator-External protocols: there are some ways in which the simulator can be examined or controlled externally.

e - Archiving protocols: people also pass archives containing whole regions (OpenSim Archives - OARs) or inventory (Inventory Archives - IARs) between OpenSimulator installations.

OpenSim presents a set of flexible protocols. It has many solutions to the communication within and out of the virtual world. This work focuses the external communication with the virtual world. A very important protocol to connect external things with the virtual world is the XML-RPC. It is better describe in section 2.3.

2.3 XML-RPC

According [27] XML-RPC is a spec and a set of implementations that allow software running on disparate operating systems, running in different environments to make procedure calls over the Internet. It's remote procedure calling using HTTP as the transport and XML as the encoding. XML-RPC is designed to be as simple as possible, while allowing complex data structures to be transmitted processed and returned.

To [25] XML-RPC is a standard for sending Procedure Calls to remote systems. It sends XML data over HTTP that remote system then handles. LSL receives XML-RPC requests and passes them to the prim specified. It may not establish this connection, but it may reply and keep two-way communication with that server.

These responses seem to be able to transport the largest amount of data out of Open Simulator or Second Life.

Both data exchange with RemoteAdmin as XML-RPC requires adjustments in server setup file. Is necessary configuring these modules in OpenSim.ini file. Once the modules are configured, it is possible to communicate using these protocols.

3 Real to Virtual Experiments

After enabling communication modules in the OpenSim.ini file and starting the virtual server, we tested the RemoteAdmin protocol. Opening a command prompt, executing RemoteAdmin.exe command with IP server address, port, password, remote administrator's command and parameters the data are send to virtual world server. A figure 1 presents command execution and the server response in the left side, and the result to virtual world viewer in the right side.

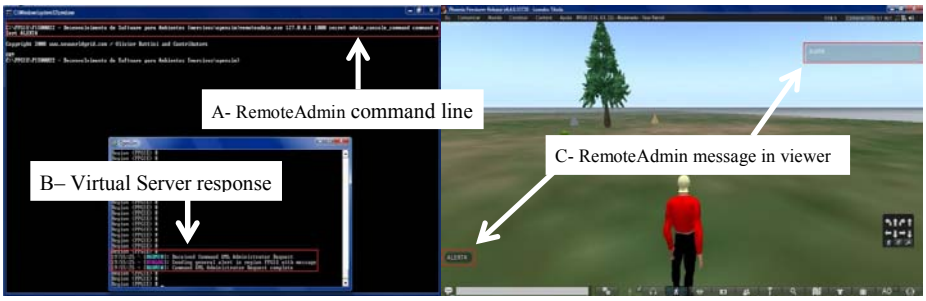


Fig. 1. Simulator-External protocols - RemoteAdmin example.

Second experiment was HTTP request to a WWW server. In virtual world the user type an integer number in a predefined channel. The LSL script sends a HTTP request to a remote server. The WWW server executes the factorial.php script returning the factorial value of the number received. The figure 2 presents the result of the request, after the avatar has touched the object.

Figure 2: HTTP request from OpenSim object.

Other experience is reverse request, from a client to the OpenSim server. To achieve successful communication is necessary import of the Apache XML-RPC into client at compilation time. We develop a Java client that pass IP server, port, channel and an integer number to OpenSim server. The channel is a link with the object and script that calculate factorial. Figure 3 exhibit client Java execution in the left side. This command connecting the server and open the channel to XML-RPC transfer messages. The white double arrow shows both clients as server execution.

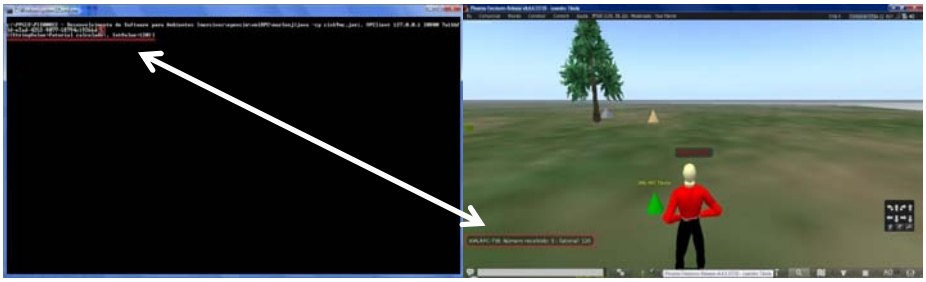


Fig. 3. External client accessing OpenSim server by XML-RPC.

All computer systems need database access. Insert, update and search of data are operations that maintain the systems status updated. So, if we want synchronize the real and virtual worlds, we need store and handle data from both worlds. The data access was developed by a research that return the e-mail to user informed. From select operation, other operations can be performed. Figure 4 shows select command result after avatar executes a remote script by a touch in the object.

Figure 4: SQL query on the remote server

4 Conclusions and Future Works

The virtual worlds is no longer science fiction and movie theme and nowadays take place in the usage everyday of technology. Adults, children, professionals, companies, institutions, and governments take advantage of virtual worlds' benefits. This paper

reported research and experiments related to real and virtual worlds communication, using the Open Simulator server. It was described the internal communication between avatars, objects and scripts in this environment; their communication protocols with external world, detailing the XML-RPC protocol. We presented several experiments involving external world data exchange through remote administration, bidirectional HTTP protocol transfer, bidirectional XML-RPC protocol with Java external client, and we performed database's operations with database information and virtual world data. To future, new experiences are planned using mobile devices, access to other remote environments, and to devices in general (Internet of Things). Data exchange with other information systems, like Learning Management System (LMS) or Enterprise resource planning (ERP) are also areas that need to be researched in more advanced way.

References

1. Wii | Assistência | Nintendo, http://www.nintendo.pt/Assistencia/Wii/Wii-301698.html
2. Firsts Steps with Kinect - Xbox.com, http://www.xbox.com/pt-BR/Kinect/GetStarted/
3. How to market your products - Second Life, http://community.secondlife.com/t5/English-Knowledge-Base/How-to-market-your-products/ta-p/700181
4. Adjei, M. T., Noble, S. M., Noble, C. H.: The influence of C2C communications in online brand communities on customer purchase behavior, Journal of the Academy of Marketing Science, Volume 38, Issue 5, pp 634-653 (2010)
5. Hodge, E., Collins, S., Giordano T.: The virtual worlds handbook: how to use second life and other 3D virtual environments, Jones and Bartlett Publishers, Sudbury (2011)
6. Rico, M., Martínez-Muñoz, G., Alaman, X., Camacho, D., Pulido, E.: A Programming Experience of High School Students in a Virtual World Platform, http://aida.ii.uam.es:8080/vleaf/es/manuales/VLEAF_IJEE.pdf (2011)
7. Kim, S.-K., Joo, Y. S., Shin, M., Han, S., Han, J.J.: Virtual world control system using sensed information and adaptation engine, In: Signal Processing: Image Communication 28, pp. 87–96 (2013)
8. Campi, A., Gottlob, G., Hoye, B.: Wormholes of Communication: Interfacing Virtual Worlds and the Real World, In: International Conference on Advanced Information Networking and Applications, pp. 2-9 (2009)
9. Guo, J., Chow, A., Gong, Z.: Virtual Wealth Realization in Virtual and Real Worlds, In: IEEE International Conference on e-Business Engineering, pp. 85-94 (2009)
10. Eno, J. D., Thompson, C.W.: Virtual and Real-World Ontology Services, In: IEEE Internet Computing, vol. 15, no. 5, pp. 46-52 (2011)
11. Chen, B, Huang, F., Lin, H., Hu, M.: VCUHK: Integrating the Real into a 3D Campus in Networked Virtual Worlds, In: International Conference on Cyberworlds, pp. 302-308 (2010)
12. Szabo, G., Veres, A., Molnar, S.: Towards Understanding the Evolution of Wars in Virtual and Real Worlds, In: Fourth International Conference on Systems and Networks Communications, pp. 153-158 (2009)
13. Back, M., Kimber, D., Rieffel, E., Dunnigan, A., Liew, B., Gattepally, S., Foote, J., Shingu, J., Vaughan, J.: The virtual chocolate factory: mixed reality industrial collaboration and control, In: International Conference on Multimedia, pp. 1505-1506 (2010)
14. Xiao-lin, W., Qiang, Y., Qi, L.: The Influence of Real World on Virtual World: An Investigation of Virtual Land Price in Second Life, In: International Conference on Management Science & Engineering, pp. 625-630 (2010)

15.Rahman, A.S.M., El Saddik, A.: HKiss: Real world based haptic interaction with virtual 3D avatars, In: International Conference on Multimedia and Expo, pp. 1-6 (2011)
16.Mishra, S., Agarwal, A., Khemka, V., Sharma, G.: Enhancing Enterprise Virtual Worlds with Real-World Event Information, In: International Conference on Computer Communications and Networks (ICCCN), pp. 1-4 (2011)
[17] Shaikh, M. Al M., Prendinger, H., Hirose, K., Mitsuru, I.: Easy Living in the Virtual World: A Noble Approach to Integrate Real World Activities to Virtual Worlds, In: International Conference on Web Intelligence and Intelligent Agent Technology, pp. 466-473 (2009)
18.Rahman, A. S. M., Hossain, S. K. A., El Saddik, A.: Bridging the Gap between Virtual and Real World by Bringing an Interpersonal Haptic Communication System in Second Life, International Symposium on Multimedia, pp. 228-235 (2010)
19.Laaki, H., Kaurila, K., Ots, K., Nuckchady, V., Belimpasakis, P.: Augmenting virtual worlds with real-life data from mobile devices, In: Virtual Reality Conference (VR), pp. 281-282 (2010)
20.Gutierrez, J.M., Oton, S., Jimenez, L., Barchino, R.: M-learning Enhancement Using 3D Worlds, International Journal of Engineering Education, Vol. 24, Issue 1, pp. 56-61 (2008)
21.Eguchi, A., Nguyen, H., Thompson, C. W.: Everything is alive: towards the future wisdom Web of things, In: World Wide Web, Vol 16, Issue 4, pp. 357-378 (2013)
22.OpenSimulator Main Page, http://opensimulator.org/wiki/Main_Page
23.Second Life Official Site - Virtual Worlds, Avatars, Free 3D Chat, http://secondlife.com/
24.Communication Protocols - OpenSim, http://opensimulator.org/wiki/Communication_Protocols
25.Second Life Wiki, http://wiki.secondlife.com/wiki/Main_Page
26.LSL Wiki HomePage, http://lslwiki.net/lslwiki/wakka.php
27.XML-RPC Home, http://xmlrpc.scripting.com/default.html

Tecnología informática aplicada a la educación de adultos mayores

Beatriz Depetris, Guillermo Feierherd, Marcela Jerez

Instituto de Desarrollo Económico e Innovación
Universidad Nacional de Tierra del Fuego, Antártida e Islas del Atlántico Sur
Hipólito Irigoyen 880 - Ushuaia - Tierra del Fuego
bdepetris@untdf.edu.ar, gfeierherd@untdf.edu.ar,
mjerez@untdf.edu.ar

Abstract. En una sociedad que envejece mejorar la calidad de vida de los Adultos Mayores (AM) es un compromiso del que las Universidades no pueden permanecer ajenas si quieren sostener su pertinencia.

Este artículo relata las experiencias de capacitación de AM en el uso de TICs, que se vienen llevando a cabo en la Sede Ushuaia de la UNPSJB (hoy UNTDF) desde 2008, y muestra como evolucionaron en función de los intereses de los participantes y la disponibilidad de recursos tecnológicos.

Las experiencias se realizaron enmarcadas en el programa UPAMI (Universidad Para Adultos Mayores Integrados), a través del cual varias Universidades Nacionales han implementado distintos cursos para AM. En el caso particular de los cursos de TICs el objetivo es disminuir la brecha digital. Las distintas experiencias han demostrado que se puede seguir aprendiendo hasta una edad avanzada, y que los nuevos aprendizajes mejoran la calidad de vida.

Los resultados se consideran altamente satisfactorios. Un grupo significativo de los participantes continúa teniendo presencia en las redes sociales.

Keywords: Brecha digital. Adultos mayores. E-inclusión.

Introducción

El objetivo del presente trabajo es describir los procesos de capacitación, las herramientas tecnológicas seleccionadas, los resultados obtenidos, las barreras encontradas y las soluciones propuestas, en los procesos de capacitación de Adultos

Mayores (AM) en el uso de TICs, llevados a cabo en la ciudad de Ushuaia durante los últimos cinco años (2008 – 2013).

De las experiencias realizadas podemos inferir que, seleccionando adecuadamente los contenidos, las herramientas informáticas y las metodologías de enseñanza para adecuarlas al grupo de alumnos, es posible integrar a los AM, con independencia de sus niveles educativos previos, a la actual sociedad del conocimiento.

Consideramos que con este modesto aporte ayudamos a disminuir la brecha digital [1], y a que este grupo etario adquiera nuevos aprendizajes que contribuyen a mejorar su calidad de vida y favorecen la autotransformación [2], así como a producir un cambio importante en la posición que este grupo ocupa en la sociedad, contribuyendo así al desarrollo social y económico de nuestro país.

El artículo está dividido en tres secciones. En la primera parte contextualizamos el tema, con una breve introducción de las características normativas del proyecto y una fundamentación de su importancia, tanto a nivel mundial como a nivel país, detallando los aspectos particulares de la Provincia de Tierra del Fuego que permiten suponer mayores beneficios al aplicar la experiencia localmente. Posteriormente se describen algunas característica de la población objetivo (los afiliados al PAMI residentes en Ushuaia). La segunda parte se refiere específicamente a los cursos desarrollados, los cambios en su contenido (consecuencia de los avances tecnológicos y los intereses de los alumnos), las barreras encontradas y las alternativas propuestas para superarlas. La tercera y última sección muestra las conclusiones obtenidas.

Contextualización

1.1 Antecedentes

En abril de 2008 fuimos convocados por las autoridades de la Regional Tierra del Fuego del PAMI (Instituto Nacional de Servicios Sociales para Jubilados y Pensionados), en nuestro carácter de representantes de la Facultad de Ingeniería de la UNPSJB, con fin de proponernos el dictado de cursos de capacitación en el área de Informática dentro del marco del programa UPAMI (Universidad para Adultos Mayores Integrados). Este es un programa establecido en un Acuerdo Marco de Cooperación entre el Consejo Interuniversitario Nacional (CIN), que nuclea voluntariamente a todas las Universidades Nacionales Argentinas, y el PAMI (Instituto Nacional de Servicios Sociales para Jubilados y Pensionados).

Un grupo pequeño de docentes de la Sede Ushuaia de la Facultad de Ingeniería, conscientes que el conocimiento y la información constituyen un intangible altamente valorado y que el acceso a ellos está cada vez más condicionado por la capacidad de utilizar las TICs, decidimos aceptar el desafío que representaba el dictado de estos cursos.

Es así que se da inicio al programa de capacitación con la firma de un convenio específico y un acta complementaria entre las autoridades de ambas instituciones. A partir de la transferencia de UNPSJB a la UNTDF, la nueva Universidad resolvió continuar con el programa.

La finalidad del proyecto es que los asistentes "adquieran habilidades y destrezas para afrontar nuevas demandas, recuperen y valoren saberes personales y sociales, y se produzca un crecimiento del diálogo inter-generacional que facilite la inserción de los adultos mayores al medio socio comunitario. Como objetivos se pretende mejorar la calidad de vida, promover el crecimiento personal y hacer efectiva la igualdad de oportunidades de ese grupo etario" .

1.2 Importancia del tema

El envejecimiento de la población mundial es, según los expertos, consecuencia de una disminución de las tasas de natalidad y un aumento de la expectativa de vida. Se trata de un cambio demográfico sin precedentes que, si bien comenzó durante el siglo pasado en los países desarrollados, afecta hoy a prácticamente todo el mundo. Al resultar poco probable que se vuelva a las elevadas tasas de natalidad existentes en el pasado, el fenómeno, que tiene consecuencias en varias esferas de la vida (económica, social, política), resulta irreversible

El informe World Population Ageing, 2009 [3], así como la actualización de algunos de sus datos al año 2012 [4], producidos ambos por la División de Población del Departamento de Asuntos Económicos y Sociales (DESA) de las Naciones Unidas, nos brinda algunos datos interesantes a nivel mundial:

- La población de 60 años o más se triplicó entre los años 1950 (200 millones) y 2000 (600 millones). Las proyecciones indican que volverá a triplicarse en el 2050 (2000 millones).
- La población de 60 años o más crece en el mundo a una tasa anual del 2,6%, en tanto que el total de la población lo hace a una mucho más modesta del 1,2%.
- A su vez la población adulta envejece. El mayor crecimiento dentro de la misma se da en los mayores de 80 años (una tasa del 4% anual).
- La tasa de envejecimiento es mayor en los países en vías de desarrollo que en los países desarrollados. El proceso que en Europa llevó dos siglos se reproduce en unas pocas décadas en varios países de América Latina.

No obstante corresponde señalar que, más allá de la cuestión cuantitativa del envejecimiento poblacional, el grupo que está desarrollando la experiencia coincide con la postura de Jesús García Minguez (García Minguez, 2009) en cuanto a que el derecho a la educación, proclamado en la mayoría de las Constituciones, no debe restringirse a una determinada edad y debe hacerse efectivo durante toda la vida del ciudadano.

1.3 Argentina

Según el informe citado [3], sobre un total de 196 naciones nuestro país ocupaba el puesto 58° en el ranking de porcentaje de población de 60 años o más, con un 14,6%, y el puesto 70° en un ranking de edad mediana, con un valor de 30,2 años.

Si bien estos datos, publicados con anterioridad al Censo Nacional de Población, Hogares y Vivienda 2010 [5], difieren levemente de los obtenidos en el mismo, lo cierto es que la población de nuestro país envejece. Un análisis de los datos de los CNPHyV de 2001 y 2010 [5] muestra que la población de 60 o más años representaba el 13,4% en 2001 y el 14,3% en 2010, en tanto que la población de 65 o más años pasó de un 9,9% a un 10,2%.

Por su parte, la proyección de población mundial incluida en el informe citado en [4], utilizando la variante media de natalidad, indica que hacia el 2040 un cuarto de la población de nuestro país tendrá 60 años o más.

Las causas de estas profundas transformaciones en la estructura de la población argentina, que disminuyen su crecimiento y modifican su estructura por edades, son las mismas que en el resto del mundo: reducción de la mortalidad infantil y de la natalidad, elección de familias más pequeñas, postergación de la llegada del primer hijo, mayor expectativa de vida, etc.

1.4 Ushuaia

Si bien en función de las características demográficas de Tierra del Fuego los datos porcentuales son similares para las dos localidades principales, nos referiremos a los correspondientes a la ciudad de Ushuaia en razón de haber desarrollado allí la experiencia.

Algunos datos de interés a partir del análisis de los datos censales de 2001 y 2010:

- Tierra del Fuego ocupa el segundo lugar cuando se toma como indicador la variación intercensal de la población total. Creció un 25,8%, siendo superada únicamente por la provincia patagónica de Santa Cruz (39,1%). En particular la ciudad de Ushuaia creció un 24,4%.
- Por su parte, la variación intercensal de la población de 60 años fue de un 81,1% y la de 65 o más de un 75,3%.

Otra característica que corresponde destacar tiene que ver con la disponibilidad de tecnología en los hogares. Una elaboración de los datos censales de 2010 permite determinar que el 79,5 de las personas de 3 o más años que vive en las viviendas particulares de Ushuaia dispone de computadora. Este porcentaje es el mayor del país. Supera ampliamente la media (47,3%), es casi un 11% más que el de la segunda jurisdicción en ese ranking (CABA, 68,6%) y más que triplica el de la jurisdicción peor posicionada (Santiago del Estero, 23,4%).

Diseño y evolución de la experiencia

1.5 Actividades preliminares

La experiencia a la que se refiere este trabajo se inició durante el segundo cuatrimestre de 2008. Durante el primer cuatrimestre de ese año se mantuvieron reuniones con directivos y profesionales de la Delegación Ushuaia del PAMI,

buscando determinar características de la población y establecer la forma en que se organizarían las actividades.

Los datos de la Tabla 1 resumen información de los afiliados a PAMI radicados en la ciudad de Ushuaia.

AFILIADOS PAMI USHUAIA

		TOTAL	VARONES		MUJERES		INDICE DE MASCULINIDAD
			PERS	%	PERS	%	
TOTAL USHUAIA		2.225	800	36,0%	1.425	64,0%	56,1
60 y más	PERS	1.791	630	35,2%	1.161	64,8%	54,3
	%	80,4%	78,8%		81,2%		
65 y más	PERS	1.479	562	38,0%	917	62,0%	61,3
	%	66,5%	70,3%		64,4%		

Fuente: Elaboración propia a partir de los datos suministrados por Delegación PAMI - Ushuaia

Tabla 1.

Otra característica significativa, señalada por los funcionarios locales del PAMI, fue que muchos de los afiliados eran personas que habían residido, hasta obtener su jubilación, en otros lugares del país. Una vez alcanzado el beneficio se habían trasladado a Tierra del Fuego acompañando a sus hijos, que a su vez lo habían hecho como parte del fuerte proceso de migración interna. Este acompañamiento respondía básicamente a dos razones: a) ayudar como abuelos en la atención de los nietos, pues es común que ambos padres estén ausentes del hogar durante muchas horas; b) la necesidad de vivir acompañados y no tener otra familia en su lugar de origen. En esos casos se produce una situación de desarraigo y de aislamiento, agravada por la necesidad de adaptarse a un clima riguroso y muy distinto de aquel al que estaban acostumbrados pues, si bien no hay información censal disponible, es posible afirmar que la mayor parte de la migración es de provincias situadas al norte de la región patagónica.

En base a la experiencia de los docentes en capacitación a otros sectores de la sociedad, la Facultad solicitó realizar los cursos con condiciones que diferían de las que el Instituto había acordado en otros lugares del país. En particular se estableció:

- la necesidad de trabajar sobre la base de una computadora por alumno, lo que, dadas las características de los laboratorios disponibles, limitaba el número de asistentes a no más de doce.
- incorporar a cada curso, además del profesor, a dos alumnos que se desempeñaran como auxiliares, a fin de brindar atención personalizada a los participantes.
- adoptar medidas para facilitar la concurrencia a fin de evitar problemas de discontinuidad, para lo que el PAMI contrató un servicio que buscaba a los alumnos en su domicilio y los reintegraba al mismo al finalizar la clase

A partir de estas primeras definiciones se realizaron las siguientes tareas:

Selección y capacitación de los alumnos que participarían de la actividad: Si bien los seleccionados tenían experiencia en el dictado de cursos, nunca lo habían hecho con una población de estas características. Por ello, convencidos que enseñar a usar TICs a AM no es lo mismo que hacerlo para niños o adultos jóvenes, fundamentalmente porque las TICs estuvieron ausentes en sus procesos educativos, en la capacitación de los futuros auxiliares se insistió en identificar las barreras que aparecerían durante el proceso (temor que la tecnología puede despertar en los AM, inseguridad sobre su capacidad para utilizarlas, problemas físicos, visuales y de motricidad propios de la edad, etc.)

Se enfatizó que la metodología de enseñanza debía ser paso a paso, con explicaciones detalladas y con un permanente acompañamiento hasta que adquirieran confianza y pudieran apropiarse de las herramientas tecnológicas. Debían tener presente que si bien la mayoría de los jóvenes aprende por autoaprendizaje (en muchos casos prueba y error) y sin temores, en este caso sería diferente.

Por otra parte, un rápido relevamiento evidenciaba que la mayoría de los alumnos que tendríamos no había completado la escuela secundaria y que en general hacía muchos años que estaban alejados de cualquier proceso educativo. En consecuencia, la paciencia era una herramienta fundamental.

Selección de los contenidos y de las actividades: Para ello se tuvo en cuenta el perfil de los alumnos y las aplicaciones que podrían despertar su interés. Una restricción importante sobre la que no fue posible introducir modificaciones (al menos al comienzo de la experiencia), fue el tiempo dedicado a cada curso, por lo que debieron plantearse en ocho (8) clases de dos (2) horas cada una, a razón de dos (2) clases semanales.

Preparación de los materiales: Se confeccionó una breve guía que se distribuyó impresa a cada uno de los participantes de los cursos.

Confección de encuestas iniciales y finales: La encuesta inicial se diseñó con el objetivo de caracterizar los grupos en cuanto a posibles conocimientos previos, posibilidad de acceder a computadora y a internet en el hogar, expectativas, etc. La encuesta final buscaba medir el nivel de satisfacción y detectar las principales dificultades encontradas por los alumnos.

Determinación de pautas que debían seguir las autoridades del PAMI para el armado de los grupos: Se elaboraron un conjunto de sugerencias buscando lograr cierta homogeneidad de los grupos en función de la edad, conocimientos previos y nivel de educación.

Adecuación del laboratorio: Se reordenó uno de los laboratorios de la Facultad buscando homogeneizar el hardware y el software que se utilizaría para los cursos. Se instaló una única versión de sistema operativo que fue configurada en forma idéntica en todas las máquinas, garantizando que lo que se proyectaba desde la computadora del profesor era exactamente lo que el alumno veía en la suya. Se ajustó la configuración de los mouses (velocidad del cursor, doble clic, etc.) y del teclado (repetición). En los que casos que era necesario se agrandó el tamaño de las letras.

1.6 Contenidos de los cursos.

La propuesta original contemplaba dos cursos. El primero de ellos, Inicial, incluía contenidos básicos para la operación de computadoras que se impartían utilizando aplicaciones de Internet (navegación, correo electrónico y chat). Estos contenidos fueron seleccionados porque posibilitaban nuevas formas de comunicación. Como ya lo mencionamos, la mayoría de los alumnos han llegado a Ushuaia después de obtener su jubilación, dejando a grandes distancias a familiares y amigos, por lo que utilizar la computadora como medio de comunicación nos parecía una tarea fundamental y motivadora para ellos. Por otra parte, el uso del navegador les daría la posibilidad de ampliar sus conocimientos (como mejorar habilidades en tareas manuales, conocer virtualmente otros lugares del mundo, leer artículos en línea, etc.) Otro objetivo importante era que lograran adquirir la capacidad para realizar trámites y pago de servicios en línea, lo que constituye un valor agregado fundamental en función de las características de los alumnos y el clima de nuestra región.

Un segundo curso incorporó conocimientos más profundos sobre administración de archivos (organización en carpetas), búsquedas avanzadas en internet, conocimientos elementales de virus y antivirus, uso de memorias externas (pendrives, memorias de cámaras fotográficas, etc.)

Este esquema de cursos se utilizó durante los años 2008 al 2010.

Inclusión de las redes sociales.

Posteriormente, atendiendo a los intereses manifestados por los alumnos y a la mayor difusión de las aplicaciones de la web 2.0, se incorporaron al segundo nivel el uso de redes sociales (Facebook y Twitter).

La contribución de las mismas para mejorar la calidad de vida de los AM ha quedado demostrada por estudios realizados en varios países. Por ejemplo, una investigación llevada a cabo en la Universidad de Arizona por Janelle Wohltmann [6] ha tenido como objetivo determinar la influencia cognitiva ejercida por el uso de una red social. Para ello la investigadora seleccionó tres grupos de 14 adultos mayores entre 68 y 91 años. Al primer grupo lo capacitó en el uso de Facebook. Sus integrantes debían interactuar solamente entre ellos y actualizar su estado al menos una vez por día. El segundo grupo fue instruido para utilizar un diario privado on-line (www.penzu.com), en el cual debían realizar diariamente al menos un registro de sus actividades. Esta aplicación no contempla interacción con otros usuarios. El tercer grupo fue utilizado como grupo de control. Al comenzar el trabajo de investigación se evaluaron variables sociales (soledad, apoyo social) y habilidades cognitivas de cada uno de los participantes. Luego de ocho semanas se volvieron a realizar las evaluaciones. Los resultados arrojaron que los integrantes del primer grupo resolvieron un 25% más de lo que hicieron al principio, mientras que en los otros dos grupos no se manifestaron cambios. Según expresa el reporte de la investigación, “los resultados preliminares ofrecen un vínculo creíble entre la conexión social y el rendimiento cognitivo.”

Otras investigaciones han mostrado un incremento en la cantidad de usuarios mayores de 50 años que utilizan las redes sociales. Por ejemplo, un reporte de Jean

Koppen [7], realizado en junio de 2010, expresa que sobre una encuesta a 1360 ciudadanos norteamericanos y 503 hispanos mayores de 50 años, aproximadamente un cuarto de los encuestados utiliza sitios web sociales, y un 23% tiene una página en Facebook que utiliza para conectarse con sus familiares y amigos. El 50% de los adultos que utilizan Facebook han incorporado la red social a través de un familiar (hijos y nietos).

Incorporación de Skype.

A partir del año 2012 se incorporó Skype (en reemplazo del Messenger), para realizar procesos de chateo. Skype es una herramienta que posibilita, básicamente en forma sincrónica, la comunicación de texto, voz y video a través de Internet. A su vez permite compartir archivos, establecer charlas virtuales, crear grupos de discusión, etc., características que tienen un uso educativo además del tradicional de comunicarse con amigos y familiares.

También permite compartir la pantalla de una computadora con otro usuario de Skype, lo que facilita mostrar presentaciones o enseñar cómo hacer algo con la computadora.

El aprendizaje de esta herramienta ha permitido que los alumnos puedan comunicarse, utilizando audio y video de calidad aceptable, con amigos y familiares que no se encuentran en nuestra ciudad. Su apropiación por los alumnos se refleja en las comunicaciones que mantienen mediante ella con el equipo docente.

Barreras encontradas y soluciones propuestas.

Las mayores dificultades mencionadas por los asistentes en las encuestas finales estaban relacionadas con la práctica de lo aprendido en el curso. Si bien muchos tenían acceso a una computadora en el hogar, señalaban que para afianzar sus conocimientos requerían utilizarla contando con alguien a quién consultar en caso de encontrar dificultades. Analizado el problema, determinamos que las soluciones posibles eran:

- Duplicar la duración de los cursos iniciales, sin introducir nuevos contenidos, y aumentando el trabajo práctico.
- Ofrecerles el uso del laboratorio de la Universidad, con la asistencia de un ayudante, dos veces por semana.

Como consecuencia se acordó con las autoridades de PAMI que, a partir del año 2010, los cursos iniciales tendrían una duración de 16 clases, lo que permitiría aumentar la práctica sobre cada tema impartido. Por otra parte, se destinó una partida para solventar las clases de consulta. Sin embargo fue imposible lograr que se facilitara el transporte para la concurrencia a estas últimas.

Estas clases de consulta, denominadas "el cyber" por los participantes de la actividad, estaban habilitadas para todos aquellos afiliados de PAMI que habían realizado algunos de los cursos previos. Es una excelente alternativa de aprendizaje, que les permite madurar los conceptos previamente aprendidos e incorporar otros de su interés. No obstante, al no haberse resuelto la cuestión del transporte, no es accesible a todos los interesados. Las características climáticas del lugar dificultan el

acceso a quienes no poseen movilidad propia o una situación económica que les permita solventar los costos de transporte público en como taxis o remises, ya que el servicio de colectivos es deficiente.

Fue por ello que a partir de este año se implementó un "cyber virtual", dos veces por semana en horarios preestablecidos. En esos horarios los auxiliares están disponibles en una cuenta premium de Skype contratada por la Universidad y los interesados en realizar consultas (que en el 90% de los casos disponen de banda ancha en sus hogares y han aprendido el uso de la herramienta en cursos previos), se comunican con ellos. La posibilidad de compartir la pantalla de una de las computadoras permite a los ayudantes guiarlos en las consultas que plantean, tal como lo harían en una situación presencial. Implementando esta opción se puede acompañar el proceso de aprendizaje de los alumnos con clases de apoyo sin la necesidad de moverse de su hogar.

Características de los participantes.

Las encuestas iniciales permitieron obtener algunas características que definen mejor la población que ha participado de los distintos cursos:

- La edad promedio es de 67 años. Poco más de un 3% supera los 80 años.
- Un 90% de los asistentes nunca habían interactuado con una computadora o, si lo habían hecho, calificaron sus conocimientos previos como “muy malos”.
- El 19% son varones y un 81% mujeres. Dado que la distribución de los afiliados del PAMI Ushuaia de 65 años o más es de un 38% de varones y un 62% de mujeres, puede establecerse un mayor interés de las mujeres por este tipo de actividades.
- Alrededor de un 40% había accedido al nivel secundario de escolaridad, un 1% a estudios superiores y el 59% restante solo al nivel primario, en unos pocos casos sin completarlo.
- A su vez, las expectativas con las que llegaban la mayoría de los alumnos estaban entre las siguientes:
  - Poder comunicarse con familiares, especialmente nietos, y amigos lejanos
  - Conocer gente de otras partes del mundo
  - Aprender a usar ese aparato que habitualmente veían usar a sus nietos
  - Aprender cosas nuevas

El porcentaje de deserción es de un 10%. De entrevistas realizadas por el personal de PAMI a los que abandonaron surge que la mayoría lo hizo por razones familiares o de salud. Fueron escasos los alumnos que desistieron del curso por considerar que el mismo superaba sus posibilidades.

Conclusiones

Desde el año 2008 y hasta la fecha se realizaron diecisiete (17) cursos iniciales y seis (6) cursos avanzados, de los que participaron aproximadamente 250 adultos mayores

Ante el asombro de los docentes, entre los que nos incluimos, nos encontramos con alumnos cuyas ganas de aprender y perseverancia superaron ampliamente nuestras expectativas. Podemos afirmar, sin duda alguna, que en más de 30 años de actividad docente pocas veces hemos visto alumnos con tanto entusiasmo. Corroboramos que "la curiosidad y el deseo de aprender no tienen límites; que el sentir y el descubrir ni se retrasa, ni desaparece con los años" [8] y por otra parte que "se puede aprender durante toda la vida, aunque el ritmo sea diferente" [9]

Consideramos que con este pequeño aporte hemos ayudado a disminuir la brecha digital, con la certeza que el acceso y uso de estas tecnologías informáticas pueden significar un cambio importante en la posición que este grupo ocupa en la sociedad

"Hay consenso en que la tecnología ofrece hoy (y potencialmente aún más) una gran oportunidad para dar respuesta a algunas de las necesidades y problemas más importantes que cotidianamente enfrentan los adultos mayores. Este grupo tiene en su haber una gran disponibilidad de tiempo libre al que podrían sacar provecho si tuvieran a su alcance estas herramientas: acceso a la formación, conectividad y financiamiento para adquirir equipos". [10]

Referencias

1. Serrano Santoyo, A. y Martínez Martínez, E.: La Brecha Digital: Mitos y Realidades. UABC. (2003) http://www.labrechadigital.org/labrecha/LaBrechaDigital_MitosyRealidades.pdf
2. Zarebski G.: Nunca es tarde para aprender. La Nación 05/03/2011 (edición impresa).
3. United Nations – Department of Economic and Social Affairs: World Population Ageing.http://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeingReport2009.pdf (2009)
4. United Nations – Department of Economic and Social Affairs: Population Ageing and Development.http://www.un.org/en/development/desa/population/publications/pdf/ageing/2012PopAgeingandDev_WallChart.pdf (2012)
5. Instituto Nacional de Estadística y Censos (INDEC). www.indec.gov.ar
6. Wohltmann, J.: Senior Citizens Who Use Facebook Have Improved Cognition. http://www.redorbit.com/news/health/1112786403/cognition-improved-senior-facebook-users-021913/ (2013),
7. Koppen, J.: Social Media and Technology Use Among Adults 50+. http://www.aarp.org/technology/social-media/info-06-2010/socmedia.html (2010)
8. García Miguez, J.: Abriendo Nuevos Campos Educativos. Hacia la Educación en Personas Mayores. Revista Historia de la Educación Latinoamericana, vol. 12, 2009, pp. 129-151. Universidad Pedagógica y Tecnológica de Colombia. Colombia (2009)
9. Ruiz Trevisan, A. y Viguera, V.: Los adultos mayores y su relación con Internet. Presentado en la mesa de Psicogerontología del II Congreso Virtual de Psiquiatría.

http://www.psiquiatria.com/bibliopsiquis/bitstream/10401/1779/1/interpsiquis_2001_1731.pdf (2001)

10. Costa Rica: Los adultos mayores y las TICs. Programa Sociedad de la Información y el Conocimiento. Hacia la Sociedad de la Información y el Conocimiento en Costa Rica, Cap. 10. http://www.prosic.ucr.ac.cr/sites/default/files/documentos/capitulo_10_4.pdf (2010)

Moderación de sesiones colaborativas a través de la virtualización de la técnica de Metaplan

Alejandro Héctor Gonzalez [1], Cristina Madoz [1], Florencia Saadi [2], Dan Hughes [2]

Calle 50 y 120 -III-LIDI- Instituto de Investigación en Informática. – La Plata Bs. As. Argentina
[1]{agonzalez, cmadoz}@lidi.info.unlp.edu.ar, [2] {florsaadi, danlaplata}@gmail.com

Abstract. Este artículo presenta la virtualización de diferentes etapas de una técnica de moderación grupal denominada "Metaplan". El formato original de la técnica se desarrolla bajo modalidad presencial y se aborda la estrategia de resolución de problemas mediante técnicas de visualización y preguntas. Se desarrolló el estudio y revisión de cada una de las etapas para proponer la virtualización del proceso. El trabajo se enmarca dentro del desarrollo de una tesina de grado en la Facultad de Informática de la UNLP. Se presenta un prototipo que permite la virtualización del Metpalan para ampliar el alcance de la técnica y facilitar el trabajo colaborativo del equipo. Se analizan los aspectos de tiempo, espacio, estilo y ritmo de cada alumno para las etapas virtuales procurando la autonomía en el desarrollo de la resolución de casos/ problemas. Se analizan los resultados obtenidos y se muestra la separación de las etapas a trabajar en forma virtual y presencial. Finalmente se desarrolla una propuesta metodológica de utilización de Metaplan en formato virtual.

Palabras claves: Metaplan, trabajo colaborativo, e-learning, groupware**.**

1. Introducción

En el mundo actual y en particular en Argentina los cambios tecnológicos y el acceso a las TIC (Tecnologías de la Información y Comunicación) están en constante desarrollo y forman parte de nuestras actividades diarias. El uso y apropiación de las tecnologías digitales genera nuevas construcciones sociales en relación a como se percibe y se entiende la tecnología. La educación no es ajena a este proceso y los estudiantes y profesores están situados en un contexto dinámico y cambiante donde se requiere la puesta en práctica de estrategias de apropiación de los medios.

La educación, y en particular la educación universitaria, transitan diferentes propuestas de revisión de las prácticas en el aula. A nivel nacional e internacional se trabaja en la creación de propuestas innovadoras que incorporen diferentes usos de tecnología digital en el ámbito educativo, permitiendo la aparición y re-significación de diferentes prácticas y modalidades de enseñanza [4][8].

Los planteamientos críticos sugieren la necesidad de reflexionar sobre: qué se enseña, cómo se enseña y cómo se evalúan los aprendizajes en los que intervienen las tecnologías digitales [12].

Se presenta como necesaria una revisión de las estrategias de enseñanza a través de un proceso de reflexión de las prácticas educativas. Este proceso debe estar

acompañado de su contextualización en el marco donde se desarrolla el proceso educativo [8].

Cabero afirma que estamos cambiando de una sociedad de la memoria a una sociedad del conocimiento, donde la inteligencia de memoria se sustituye por una inteligencia distribuida apoyada en los diferentes instrumentos tecnológicos [4]. Esto trae como consecuencia la aparición de un nuevo tipo de inteligencia denominada ambiental que existirá como consecuencia de la interacción con las distintas TIC.

Entre las estrategias de incorporación de TIC al proceso educativo se puede encontrar el trabajo grupal en línea, donde participan varias personas a fin de llegar a mayor diversidad de conceptos y criterios por medio de herramientas digitales que favorezcan su interacción [15].

Existen diferentes técnicas de aprendizaje que promueven la distribución y comunicación entre los participantes de un grupo. La técnica de Metaplan se puede considerar como una metodología de moderación grupal que facilita la obtención de resultados por medio de visualización y preguntas. Se la puede implementar en diferentes campos de acción como: planificación, solución de problemas, toma de decisiones participativas, diagnostico de necesidades, evaluaciones grupales y retroalimentación, procesos de enseñanza y aprendizaje, debates y talleres, entre otros [5][6] [11].

Esta técnica en su formato original se aplica en forma totalmente presencial. Resulta atrayente aportar la incorporación de tecnología digital a esta técnica para ampliar el alcance, integrar más participantes en diferentes espacios y tiempos. Se propone virtualizar ciertas etapas, promoviendo la interacción grupal en la elaboración de ideas y conocimientos [14].

2. Marco teórico

Se trabaja el concepto de trabajo colaborativo presencial y virtual basándose en el enfoque de la Cognición Distribuida que intenta entender la organización de los sistemas cognitivos de las personas y entornos. Los límites de lo cognitivo pueden ir más allá de lo individual para abarcar fenómenos que suelen aparecer en las interacciones entre las personas y los ambientes en que estas se mueven [1] [2] [3].

La cognición distribuida es una rama de la ciencia cognitiva que propone que el conocimiento humano y la cognición no se limitan solamente a las personas. Esta perspectiva permite analizar la comunicación entre los artefactos y su contexto [18].

Los procesos cognitivos pueden ser distribuidos entre los miembros de un grupo social o comunidad. Se distribuyen en el sentido de que el funcionamiento del sistema cognitivo incluye la coordinación entre los componentes internos y externos (de carácter material o ambiental) de su estructura.

Los procesos también pueden ser distribuidos a través de un determinado tiempo, de tal manera que los acontecimientos anteriores puede transformar la naturaleza de los eventos relacionados en un ecosistema cognitivo

Desde el aspecto informático el denominado “groupware” se lo define integrando el software y la parte humana. En el groupware se consideran los problemas técnicos, las implicaciones organizacionales y sociales de introducir TIC. Se debe trabajar el

desarrollo del trabajo grupal a través de medios digitales procurando la adaptación en apoyo al objetivo del grupo y al proceso utilizado. Es necesario que la evolución del sistema humano y el tecnológico se encuentren equilibrados para no perder de vista las implicaciones sociales y poder crear nuevas estructuras organizacionales y roles [14] [16].

El contexto que propician las TIC ofrece a los docentes un abanico de herramientas y deben estar preparados para utilizarlas adecuadamente. Un término asociado a la relación docentes-TIC se denomina “TAC” (Tecnologías para el aprendizaje y el conocimiento) [17]. Juana Sancho reflexiona sobre el uso de las TIC y propone un juego de palabras denominando TIC/TAC

Las TAC tratan de orientar las TIC hacia usos más formativos, tanto para el estudiante como para el profesor, con el objetivo de aprender más y mejor. Se trata de incidir especialmente en la metodología, en los usos de la tecnología y no únicamente en asegurar el dominio de una serie de herramientas informáticas. Se trata en definitiva de conocer y de explorar los posibles usos didácticos que las TIC tienen para el aprendizaje y la docencia. Las TAC van más allá de aprender meramente a usar las TIC y apuestan por explorar estas herramientas tecnológicas al servicio del aprendizaje y de la adquisición de conocimiento [13].

Este cambio de TIC a TAC va de la mano de la formación docente y la organización del sistema de enseñanza en general y el desarrollo de la práctica en las clases en particular.

Cuando un material de estudio es elaborado teniendo en cuenta las posibilidades comunicacionales del lenguaje en que está armado el mensaje y del medio a través del cual se lo ofrece puede lograrse en esa mediación una mayor relación de aprendizaje [11]. En este tema se debe destacar cómo las TIC pueden posibilitar procesos de cognición distribuida masiva que resultaría difícil de organizar de forma analógica.

Con referencia al proceso de trabajo grupal puede ser retomado desde el espacio del taller en el aula presencial. En este contexto las personas en interacción con otras suelen enriquecerse de nuevas opiniones y abordar nuevas conclusiones de manera de ampliar el conocimiento sobre algún tema en particular.

La técnica de Metaplan presenta una vía posible para conseguir motivación intrínseca por aprender y asegurar aprendizajes que favorezcan la interacción: alumno – conceptos/materiales de estudio – formador - otros alumnos, en pos de un determinado objetivo de aprendizaje [6].

El Metaplan está divido en etapas. Cada etapa fue analizada para ver la posibilidad de ser implementada de forma virtual. Surge la posibilidad de pensar en una técnica de Metaplan que se apoye en las TIC facilitando el cambio del rol docente hacia un rol de moderador y facilitador del aprendizaje y donde siga teniendo lugar el acto de enseñar. Esta propuesta puede aportar al desarrollo de prácticas docentes del tipo TIC/TAC.

2.1 Técnica de Metaplan

La técnica de Metaplan es una de las posibles técnicas para propiciar el trabajo colaborativo y puede verse como un conjunto de "herramientas de comunicación" para ser usadas en grupos que buscan ideas, soluciones para sus problemas, desarrollo

de opiniones, acuerdos, formulación de objetivos, recomendaciones y planes de acción.

Cisnado Torres indica que la técnica de Metaplan fue ideada por Eberhand Schelle en Alemania, donde el instrumento pedagógico fundamental lo constituye una "situación interaccional" que a partir de una pregunta o de una tesis presentada por el formador se provocan contestaciones simultáneas y visibles por parte de todos los participantes. Se puede mantener un tono de atención y tensión durante el proceso generado por el interés de comprobar si las otras contestaciones confirman la propia idea, si se oponen o si complementan el propio conocimiento sobre el tema tratado [5][6] .

El Metaplan se visualiza como una herramienta de trabajo aplicable en capacitaciones, talleres y reuniones. Esta metodología agrega un acompañamiento en forma de "moderador". La meta de un buen moderador es mantener vivo el interés del alumno por aprender, provocando su participación activa, animándole a hacer consultas y a presentar dudas. La interacción se presenta en apoyo del proceso de aprendizaje [10].

Desde el punto de vista psicológico el aprendizaje interaccional potencia y aprovecha la motivación intrínseca del propio proceso formativo a favor del aprendizaje que se pretende conseguir.

Existen dos tipos de motivación diferenciadas por psicólogos, la extrínseca y la intrínseca [1][9]. La primera viene por una causa externa al tema o curso en la que está involucrado el participante, y es generada por un deseo interno de evitar algo negativo o conseguir una mejora en algún aspecto de su vida. La segunda se evidencia cuando el individuo realiza una actividad por el simple placer de realizarla sin que nadie de manera obvia le de algún incentivo externo. Un hobby es un ejemplo típico, así como la sensación de placer, la autosuperación o la sensación de éxito.

Cuando se espera un beneficio por tomar cierto curso, por ejemplo un ascenso laboral o un mayor salario, se crea la tendencia a que las personas participen con una motivación extrínseca. En contextos laborales no siempre será posible contar con elementos de motivación extrínseca en cada una de las acciones formativas propuestas. En cambio cuando conseguimos el mismo deseo de participar por medio del desarrollo en el aprendizaje, cuando se logra que las personas disfruten capacitarse por el solo hecho de hacerlo, sin pensar en posteriores beneficios ajenos, se está generando motivación intrínseca.

Para un docente es fundamental contar con alumnos animados por aprender, lo cual tendrá consecuencias tanto para el formador como para el diseñador de la formación. Se procura de esta manera que los participantes:

- Desplieguen una actividad propia.
- Influyan ellos mismos en el desarrollo de las cosas.
- Hagan algo conjuntamente con otros.
- Dominen un problema nuevo.
- Tengan vivencias inmediatas de éxito.
- Conserven, a posteriori, un recuerdo satisfactorio de que todo ello ha costado un poco de esfuerzo.
- Se comprometan con algo.

El aprendizaje interaccional elimina la tutela pedagógica como modo de funcionar, lo cual implica que el participante desarrolle mayor motivación intrínseca [7].

La tutela pedagógica, genera un sentimiento inconsciente e inevitable en la persona de estar obligado a aprender por alguien que tiene un mayor conocimiento sobre la materia que se aprende. Aunque ese tutor pueda ser simplemente un libro, una persona o un tutor virtual. Las personas tienen este mecanismo arraigado en la memoria desde la infancia y cada vez que nos ponemos en situación de aprender, aparece esta sensación de necesitar que se nos guíe. Se suele asociar aprender con esfuerzo, con utilidad a medio o largo plazo, con dificultad, con competición, con situaciones en las que predomina la dialéctica de la inferioridad y el dominio.

El aprendizaje interaccional intenta lograr en el participante las ideas de:

- Es interesante para mí o de por sí.
- Puedo aprender y me importa aprender.
- Es útil, funcionará.
- Puedo hacerlo junto a otras personas que comparten realidad conmigo.

¿Cómo se desarrolla el aprendizaje interaccional en el Metaplan?. Se trabaja con un moderador, su función principal es la de ayudar a mejorar el entendimiento mutuo. Su objetivo es el de ofrecer al grupo las técnicas de comunicación necesarias, en el momento preciso para que los participantes puedan encontrar las soluciones efectivamente. El mismo comienza haciendo preguntas a los asistentes. [6]

Una vez que el moderador reúne las opiniones de los participantes, las agrupa por su similitud. Para cada idea nueva que no encuentre semejanza con las ya expuestas, se crea una nueva nube denominada “nube de ideas”, en caso contrario se agrupa con la que guarde parecido. El moderador diagrama la nube de ideas, por cada nube se genera un subtema que el moderador distribuye a los participantes de la denominada “session” de Metaplan.

El moderador es quien decide la distribución de los subgrupos y subtemas entre ellos [6]. Luego por cada grupo se arma la “lista de recomendaciones”, que resulta ser un plan de acciones que están en espera de ser aprobados y hace referencia a los temas, deseos y acciones planteados por los grupos. Estos elementos se anotan en la lista y se los destaca por orden de importancia, de esta manera quedan registrados los puntos sobre los que se debe tomar acción.

Finalmente el grupo completo realiza el debate y se genera la “lista de acciones” que refiere a las actividades que se pueden desarrollar. A cada acción a tomar se le asigna un responsable y un grupo de personas encargadas a desarrollar la acción.

El proceso completo de Metaplan sería:

- Planteo del tema central
- Aporte anónimo de cada participante sobre el tema central
- División de opiniones en subtemas-> Nube de Ideas
- División de participantes en subgrupos asignando subtema/s
- Debate de cada subgrupo de el/los subtemas asignado/s: exponer opinión y/o rankear alguna ya expuesta
- Diseño de lista de recomendaciones de cada subgrupo
- Cada subgrupo expone la lista de recomendaciones al grupo completo
- Discusión del grupo completo sobre o expuesto por los subgrupos

- Conclusión y resumen. Lista de acciones.

3. Desarrollo del prototipo

Luego de analizar el Metaplan se puedo observar que la virtualización de ciertas etapas de la técnica puede:

- Ampliar el alcance de la capacitación sobre la técnica.
- Facilitar el proceso de enseñanza y aprendizaje de la metodología, considerando aspectos de tiempo, espacio, estilo y ritmo de aprendizaje de los alumnos, promoviendo así su autonomía en este proceso.
- Aprovechar las ventajas presentes en las TIC para la facilitar el desarrollo de los talleres.

Se desarrolló un prototipo para la administración y desarrollo de espacios de trabajo vía Web con la técnica grupal de Metaplan incorporando la virtualización de las etapas necesarias que permitan llevar adelante cada tarea.

Se trabajó con un modelo de desarrollo de software denominado "prototipado evolutivo" para poder ir testeando y modificando las fases de desarrollo.

El prototipo logrado permite virtualizar las etapas de aporte anónimo de los participantes sobre el tema central. El moderador puede realizar la división de opiniones en subtemas generando las "nubes de ideas". Se pueden generar espacios de discusión y posterior planteo de la lista de recomendaciones de los subgrupos para que las personas que no puedan estar presentes en todas las sesiones del Metaplan se involucren en el desarrollo de un taller que utilice esta metodología de enseñanza.

El prototipo presenta una interfaz con tres plantillas, una por cada perfil de usuario: Administrador, Moderador y Participante.

La plantilla del *Administrador* permite realizar un control sobre los datos del sistema, incluyendo gestión de usuarios y accesos a las diferentes secciones del sitio, gestión de cursos e inscripciones.

La segunda plantilla es la interfaz de los *Moderadores*, permite gestionar los espacios de trabajo virtuales, realizar el seguimiento de la interacción de los participantes tanto en la construcción de la nube de ideas como en los foros de discusión. Presenta diferentes secciones para la gestión de grupos de participantes, temas en los que se divide el curso, y administración de los foros de discusión en los que discutirán esos grupos.

La tercer plantilla es la interfaz de los *Participantes*, medio por el que los alumnos pueden inscribirse, opinar sobre el tema central planteado, interactuar con los integrantes de su grupo por Chat (herramienta de comunicación sincrónica) y/o por los foros de discusión (asincrónicos) creados para cada tema asociado al grupo.

3.1 Requerimientos técnicos para el prototipo

La aplicación fue diseñada para poder funcionar en un ambiente utilizando software libre. Requiere que ciertas dependencias se instalen previo a su ejecución. Las mismas son:

- MySQL 5 o superior

- PHP 5.2.4 o superior
- Symfony 1.4
- Plugins de Simony a instalar: sfPropelPlugin, sfJQueryUIPlugin, sfProtoculousPlugin, sfJqueryReloadedPlugin, sfFormExtraPlugin, sfTCPDFPlugin, Servidor de correo
- Se utiliza el producto Eclipse como IDE (Entorno de Desarrollo Integrado) programando en lenguaje Php5 con un template Symfony.

En el modelo de datos inicial para el prototipo de la aplicación se identifican los siguientes elementos:

- Session: para modelar los cursos Metaplan.
- Debate: opiniones, comentarios y nubes de ideas para integrar opiniones.
- Temas: conceptos que harán que interactúen y debatan los participantes.
- Interacción: salas de chat y foros para debatir en forma virtual entre los participantes.

3.2 Funcionamiento del prototipo

El moderador es el encargado de distribuir la información de la session, administra los grupos de participantes y los subtemas de la session. Los moderadores son quienes se encargan de organizar los temas tratados en los cursos, mediar en las discusiones, agrupar los usuarios en grupos y administrar temas a tratar por los grupos. El prototipo provee foros de discusión y los moderadores son quienes se encargan tanto de la administración como la de moderación de los foros.

Los *participantes* pueden administrar la información referente a los subtemas que el moderador le haya asignado a su grupo. Se inscriben a los cursos, exponen opiniones, interactúan con sus compañeros de grupo y plantean posibles soluciones al tema planteado en la lista de recomendaciones.

Los *administradores del sitio* son los que tratan la información referente a la administración de usuarios, menúes y temas referentes a la configuración del sitio Web. Los Administradores gestionan como se puede observar en la figura 1, los tipos de perfil de usuario, menús y sus accesos.

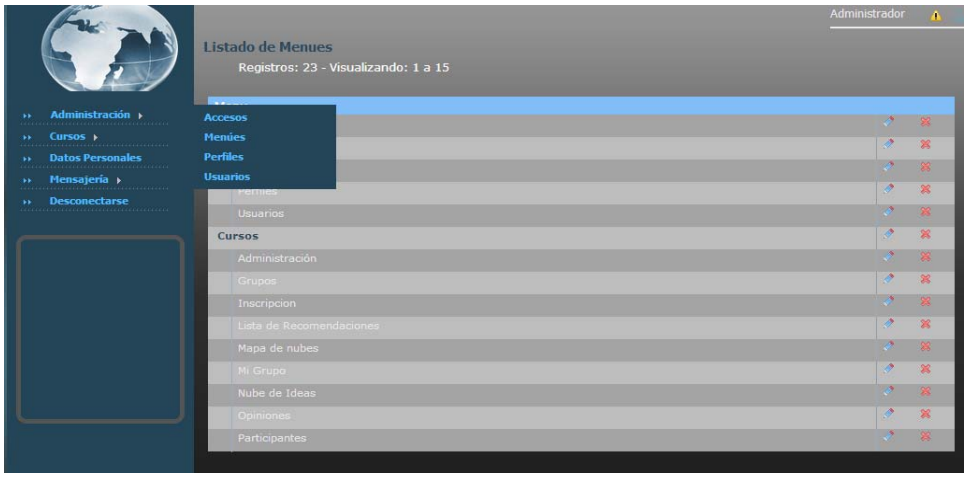


Fig1. Administración de sesiones de Metaplan

3.3 Pasos a seguir en el Metaplan virtual

Cada espacio de trabajo virtual dispone de un tema central que será el motivo de discusiones y opiniones, un rango de fechas que representan el periodo de validez del dictado del curso, y una persona que actuara de mediador entre los temas y los grupos. En la sesión inicial se plantea el "Tema central" sobre el que posteriormente los participantes opinaran anónimamente y debatirán en forma general para llegar a la "Nube de ideas", que sirve como un Mapa que reunirá las principales opiniones que se transformaran luego en los subtemas a tratar en las etapas subsiguientes. El tema central es presentado en el primer encuentro presencial del metaplan.

Una vez que cada participante expone sus opiniones, administradas en nube de ideas, las publicará de forma que el moderador pueda ir visualizándolas en el mapa de nubes. En esta etapa, los participantes del curso y el moderador se reunirán nuevamente en forma presencial y juntos armaran el mapa de nubes final del tema central.

El moderador puede realizar varias acciones, pero ninguna de estas será sin el consentimiento de los participantes del curso, debido a que el Moderador es quien los acompaña en ese proceso. Para esta labor el Moderador puede editar una nube en particular, (cambiando el titulo de la nube o eliminando una opinión de la misma), intercambiar opiniones entre las diferentes nubes expuestas y/o eliminar una nube.

El moderador al publicar el mapa de nubes producirá el resultado final de esta etapa, que serán las nubes de ideas que queden como elegidas, sus títulos serán los subtemas que debatirán los grupos de participantes posteriormente. En este proceso de publicación de temas, la session de Metaplan pasará del estado "inicializada" a "temas_publicados", con los subtemas planteados solo falta que la totalidad de los participantes se dividan de forma tal que puedan tratar los subtemas.

El proceso final de toma de decisiones y acuerdos se desarrolla en forma presencial.

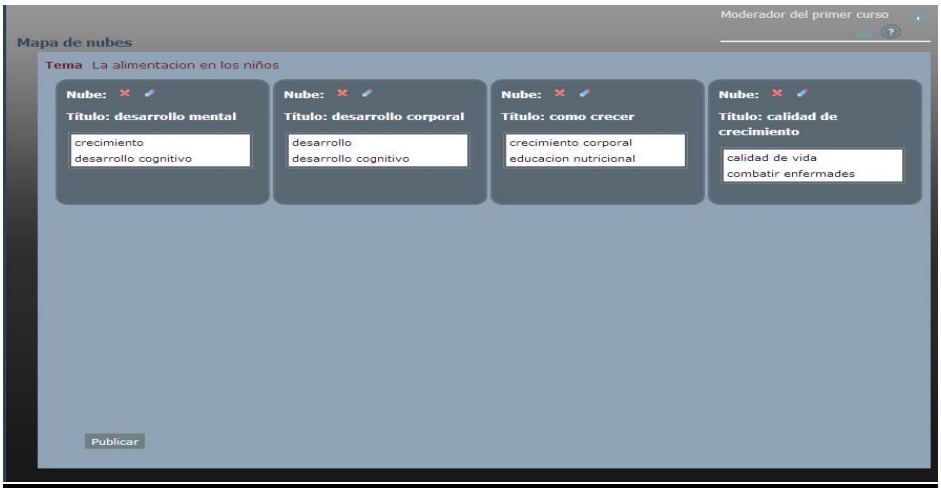


Fig 2. Nubes de ideas

4. Conclusiones y trabajo futuro

Se realizó el análisis y descripción de la técnica de Metaplan. Se trabajó en la separación de las etapas determinado cuales podrían ser trabajadas de manera virtual.

Se construyó un prototipo evolutivo para presentar las tres interfaces esenciales: administrador, moderador y alumno.

Se desarrolló la idea inicial para poder compartir y visualizar las nubes de ideas.

El análisis de la técnica de Metaplan permitió analizar cada etapa y proponer etapas virtuales. La generación de las nubes de ideas es una etapa que permite trabajar en forma asincrónica y pretende mejorar los tiempos entre cada sesión de Metaplan. Se pudo relacionar los procesos en cada etapa atendiendo al principio de cognición distribuida.

Se abordó el concepto de TIC transformados a TAC donde los docentes puedan apropiarse de la herramienta a través del trabajo en el taller y se focalicen en las TIC como medios para el aprendizaje y el conocimiento. La relación docente-conceptos-alumnos se fortalece a través del trabajo integrado de todos los participantes tanto en las instancias presenciales como virtuales. El cambio en el rol docente es fundamental para el buen desarrollo de la técnica.

Como posibles extensiones de este trabajo se puede pensar en integrar una herramienta sincrónica para mejorar el armado visual de las nubes de ideas. También es necesario lograr un seguimiento de las acciones de cada participante a fin de contar con estadísticas que ayuden a ver la colaboración y el uso de la aplicación.

A futuro se espera mejorar la implementación sobre la retroalimentación de una experiencia piloto y poder ofrecer el acceso a la herramienta desde un sitio Web.

Bibliografía

1. Ainscow, M. & West, M. (2006). *Improving urban schools. Leadership and collaboration.* Maidenhead: Open University Press.
2. Avila Patrica M., Bosco Martha D. (2001) . "Virtual environment for learning a new experience. Abstract ID: 1510. Trabajo presentado en el " 20th. International Council for Open and Distance Education". Düsseldorf, Germany.
3. Barberá Elena. (2008). "Calidad de la enseñanza 2.0. Educational quality 2.0" . RED: Revista de Educación a Distancia, ISSN 1578-7680, Nº. Extra 7. España.
4. Cabero Almenara J., María del Carmen Llorente Cejudo (2007). "Propuestas de colaboración en educación a distancia y tecnologías para el aprendizaje". Edutec. Revista Electrónica de Tecnología Educativa. Núm. 23.
5. Cisnado Torres Xiomara (2008). "Metaplan, una metodología de diagnostico y moderación grupal". Centro de capacitación. Contraloría general de la república. Costa Rica http://jaguar.cgr.go.cr/content/dav/jaguar/documentos/capacitacion/web_centro /Metaplan/Metaplan.htm
6. Cisnado Torres Xiomara (2007). Virtualización de la Enseñanza-Aprendizaje de METAPLAN, www.infodesarrollo.ec/component/docman/doc_download/132-virtualizacion-de-la- ensenanza-de-aprendizaje-de-Metaplan.html
7. Delgado Fernández M., Solano González A. (2009). "Estrategias didácticas creativas en entornos virtuales para el aprendizaje". Revista: Actualidades Investigativas en Educación. Volumen 9, Número 2 pp. 1-21.
8. Diaz Barriga F: (2011). "La innovación en la enseñanza soportada en TIC. Una mirada al futuro desde las condiciones actuales". VII Foro Latinoamericano de Educación / Experiencias y aplicaciones en el aula. Aprender y enseñar con nuevas tecnologías
9. Dror, I. E. & Harnad, S. (2008). Offloading Cognition onto Cognitive Technology. Cognition Distributed: How Cognitive Technology Extends Our Minds (pp 1–23). Amsterdam: John Benjamins Publishing.
10. EPISE, Metaplan (2005) Sesiones formativas y reuniones de trabajo más efectivas, http://www.episе.com/episecms/galeria/documentos/Metaplan_21_ene_08.pdf
11. Hanusyk Klaus. (2010). "Introducción al Método de Moderación". Disponible en la página http://www.klaushanusyk.com/ visitada en Marzo 2011.
12. Litwin E., Maggio M, Lipsman M. (2004)."Tecnologías en las aulas. Las nuevas tecnologías en las prácticas de enseñanza. Casos para el análisis". Amarrortu editores. Buenos Aires-Madrid.
13. Lozano, Roser. "Las 'TIC/TAC': de las tecnologías de la información y comunicación a las tecnologías del aprendizaje y del conocimiento". Anuario ThinkEPI, 2011, v. 5,
14. Madoz C., Gonzalez A. Saadi M., Hughes D. (2010). "Virtualización sobre un entorno de Enseñanza y Aprendizaje de métodos de trabajo colaborativo". Presentado en el TEyET, Calafate. Santa cruz. Argentina.
15. Prendes Espinosa, M. P., Martínez. (2006). "Actividades individuales versus actividades colaborativas", en E-actividades: un referente básico para la formación en Internet, ISBN 84-665-4768-1, pags. 183-202.
16. Rama, J., & Bishop, J. (2006). "A survey and comparison of cscw groupware applications". Annual research conference of the South African. Somerset West, South Africa.
17. Sancho Gil, Juana M.(2008). "De TIC a TAC, el difícil cambio de una vocal. Revista Investigación en la escuela". Núm: 64, Pág.: 19-30. Biblioteca: DIE. Universidad de Barcelona. Consultado en Abril de 2012 desde http://www.ub.edu/esbrina/docs/proj-tic/tic_a_tac.pdf
18. Solomon, G. (2005) "Distributed Cognitions. Psychological and educational considerations". Cambridge University Press.

Taxonomía de Mecanismos de Awareness

Alexander Herrera[1], Darío Rodríguez[2], Ramón García-Martínez[2]

1. Programa de Maestría en Ingeniería de Sistemas de Información. UTN-FRBA
2. Laboratorio de Investigación y Desarrollo en Espacios Virtuales de Trabajo
Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús.
drodrigu@unla.edu.ar, rgm1960@yahoo.com

Resumen. En el contexto de los sistemas de trabajo colaborativo mediado por tecnología, un grupo puede ser visto como un conjunto de individuos que interactúan directamente o por medio de artefactos compartidos y que se perciben a si mismos como un grupo. En gran parte, estas percepciones se logran a través de mecanismos de awareness. En este trabajo se presenta una comparación de los mecanismos de awareness estudiados, y se propone una Taxonomía de este tipo de mecanismos.

Palabras Clave: mecanismos de awareness, taxonomía.

1. Introducción

El término “awareness” se define en ingles como el sustantivo de “estar al tanto”, “consiente”, “tener conocimiento” o “estar informado” [Harper Collins, 2011; Cuyas, 1982]. En el contexto de los sistemas de trabajo colaborativo mediado por tecnología (Computer Suported Colaborative Work - CSCW) un grupo puede ser visto como un conjunto de individuos que interactúan directamente o por medio de artefactos compartidos y que se perciben a si mismos como un grupo. En gran parte, estas percepciones se logran a través de mecanismos de awareness.

Dourish y Belotti [1992] definen el awareness como el entendimiento de las actividades de otros, que proporciona un contexto para la propia actividad. Gutwin y Greenberg [2002] ofrecen una definición mas especifica al definirlo como el conocimiento, hasta el detalle, de las actividades de otras personas que se requiere para que una persona (el conocedor) pueda coordinar y completar su parte de una tarea de grupo. En el marco de un grupo de trabajo mediado por tecnología, el awareness entre los integrantes del grupo se mantiene mediante el acceso a información (por parte de todos los miembros del grupo) sobre cada uno de sus miembros; como por ejemplo: la ubicación de otros participantes en el espacio compartido (¿donde están trabajando?), sus acciones (¿qué están haciendo?), la historia de interacción (¿lo que lo han hecho?), y sus intenciones (¿qué van a hacer?).

En este trabajo se presenta el estado de la cuestión sobre los mecanismos awareness (sección 2), se introducen conceptos que permiten su caracterización (sección 3), se presenta una comparación de los mecanismos de awareness estudiados (sección 4), se propone una taxonomía de este tipo de mecanismos (sección 5), y se

formulan algunas conclusiones preliminares indicando futuras líneas de trabajo (sección 6).

2. Estado de la Cuestión

En [Gutwing et al., 1995] se describe una estructura que enmarca el awareness de espacio de trabajo en un contexto de requerimientos de awareness para aprendizaje colaborativo y presenta una forma de organizar situaciones de colaboración en términos de tareas y separación de vistas, e introduce diferentes tipos de componentes que apoyan el mantenimiento del awareness en espacios de trabajo.

En [Carroll et al., 2002] se sugieren estrategias de diseño para sistemas de notificación que den un mejor apoyo a actividades colaborativas con el Awareness de actividad en el cual se hace hincapié en la importancia de los factores contextuales de la actividad como la planificación y la coordinación. Se ha desarrollado un enfoque alternativo en el cual la gestión de datos de actividades da apoyo en el awareness de actividad, proveyendo líneas de tiempo para la visualización de datos y el acceso a los datos. Los elementos comunes de coordinación (plazos limite, las comunicaciones y las versiones del documento) son omnipresentes en todos los mecanismos de notificación. La colocación de estos elementos en líneas de tiempo proporciona un resumen significativo del progreso del proyecto.

En [Collazos et al., 2006] se propone un ambiente multi-agente de enseñanza/aprendizaje (ALLEGRO) en el cual se describen actividades colaborativas, awareness para este ambiente y una arquitectura blackboard. El modelo de awareness propuesto permite promover la concientización, comunicación, colaboración y coordinación en el ambiente de aprendizaje colaborativo mediado por tecnologia (Computer Suported Colaborative Learning – CSCL). Este modelo permite a los aprendices tener una percepción de lo que los demás hacen dentro del CSCL con el propósito que evaluen que les sirve para su entorno y actividades de trabajo. Además el awareness del CSCL permite al aprendiz tomar un papel activo en su proceso formativo a través de actividades que le permiten exponer e intercambiar ideas y opiniones con los demás integrantes, convirtiendo de esta forma el aula virtual en un foro abierto a la reflexión y al contraste critico de pareceres y opiniones.

En [Gutwing y Greenberg, 2002] se presenta un marco de trabajo orientado a grupos pequeños en áreas de trabajo de tamaño medio, donde es mas probable que los participantes estén interesados en mantener awareness con todos los miembros del grupo, El marco describe tres aspectos de awareness en espacios de trabajo: los elementos que lo componen, los mecanismos utilizados para sus mantenimiento y sus usos en la colaboración. La estructura del marco de trabajo puede ser usada para describir otros tipos de awareness que afectan el trabajo en grupo de forma distribuida. El ciclo de percepción-acción es un modelo general que puede ser usado para explicar como las personas realizan seguimiento de una amplia variedad de información en una situación de colaboración.

En [Collazos et al., 2002] se presenta un tipo de awareness llamado SKA (Shared Knowledge Awareness) para aumentar y mantener el conocimiento compartido en un grupo. Este tipo de awareness puede ser proporcionado de forma grafica representado

en el indicador de conocimiento compartido (SKI), que estima el conocimiento compartido a través de acciones y mensajes de los miembros del grupo. Los autores citados creen que el SKA puede tener un impacto positivo en las actividades meta cognitivas ayudando en la construcción y mantenimiento de los conocimientos compartidos. El mecanismo que da soporte a este tipo de awareness es construido con un agente inteligente y dos módulos analizadores.

En [Fuchs et al., 1995] presenta un modelo de distribución de eventos para un entorno de trabajo cooperativo basado en equipos. El modelo propuesto proporciona información sobre las actividades actuales y pasadas de los usuarios que colaboran, basados en la semántica y las relaciones contextuales de los artefactos compartidos y contribuye a aumentar el awareness sobre el actual estado de cosas sin sobrecargar al usuario con información adicional. El sistema implementa un entorno sencillo para la coordinación del trabajo distribuido y permite el apoyo del awareness compartido de los usuarios mediante la aplicación del modelo de eventos y visualización de la información del evento usando la metáfora del escritorio.

3. Caracterización del Awareness

En esta sección se definen las siguientes características del awareness: arquitectura auxiliar (seccion 3.1), modo de awareness (seccion 3.2), propagación del mensaje (seccion 3.3), tipos de aplicación (seccion 3.4), tipos de presentación (seccion 3.5) y métodos de presentación(seccion 3.6).

3.1. Arquitectura auxiliar

Esta constituida por las partes de los mecanismos de awareness que cumplen con un rol especifico dentro del sistema. Esta arquitectura puede ser tradicional como módulos de administración de información que se encargan de tareas básicas o módulos un poco mas avanzados que se encargan de clasificar o analizar. También se pueden implementar sistemas inteligentes que generen nueva información de tipo awareness, esto, con el fin de ayudar en el suministro de información contextual que de soporte al tipo de awareness en el que se este trabajando.

3.2. Modo de Awareness

Es el conjunto de eventos que permite una descripción del estado de situaciones de cooperación y permite proporcionar información para apoyar cada uno de los diferentes modos de awareness presentados a continuacion:

Sincrono Acoplado:	lo que esta actualmente pasando en el ambito de trabajo.
Asíncrono Acoplado:	lo que ha cambiado en el ambito de trabajo desde el ultimo acceso.
Sincrono Desacoplado:	lo que ocurre actualmente en cualquier otro lugar de importancia.

Asíncrono Desacoplado: Cualquier cosa de interés que haya ocurrido hace poco en otro lugar.

El awareness sincrónico se ocupa de los eventos, que en la actualidad se están produciendo, mientras que el awareness asíncrono considera eventos que han ocurrido en algún momento en el pasado. El apoyo a este último modo tiene que ser dado por una interpretación que resume toda una secuencia de acontecimientos, que han ocurrido en el íntermedio. El awareness sincrónico debe estar apoyado por un reflejo inmediato del trabajo en curso en la interfaz gráfica de usuario del sistema. Se distinguen según el interés del usuario entre el awareness acoplado y desacoplado. El awareness acoplado denota el tipo de información general, que está estrechamente relacionado con la ocupación actual del usuario. Un ejemplo de este tipo de orientación es el conocimiento de un usuario que desea editar un documento determinado, que dicho documento esta siendo leido actualmente por otra persona. Con el awareness acoplado asíncrono son las situaciones, cuando un usuario está trabajando en un determinado objeto y es informado de cambios, que le pasaron a este objeto en el pasado durante un período de ausencia. El awareness desacoplado se aplica en situaciones donde la información sobre los eventos necesita ser dada independiente del actual enfoque de trabajo del usuario. A modo de ejemplo para el awareness asíncrono desacoplado se puede considerar una situación en la que el administrador del flujo de trabajo envía un objeto (ej: hoja de cálculo, carpeta de documentos) en el que hay que trabajar, a alguien que esta en vacaciones. Si hay una fecha límite, entonces puede ser muy importante notificar al iniciador del flujo de trabajo de esto, aunque en el momento este ocupado con otra cosa.

3.3. Propagación del Mensaje

La información que es de tipo awareness tiene que ser enviada o mostrada a las personas que sean parte del CSCW, este información puede ser enviada a todas las personas del grupo (multi - destino), a una persona del grupo (destino individual), estar basado en los intereses del usuario en situaciones de trabajo (relación usuario - objeto), o de acuerdo a algún nivel jerárquico que exista en el CSCW.

3.4. Tipos de aplicación

En [Ellis et al., 1991] se identifica la manera de descomponer los sistemas colaborativos a través de una matriz espacio-temporal, de acuerdo a esta matriz se pueden identificar los siguientes tipos de aplicaciones:

- Interacción Cara a Cara: Implica el mismo tiempo y el mismo lugar, puede dividirse en varias categorías; pantalla compartida para explicaciones, entornos de conversación y tormentas de ideas
- Interacción Asíncrona Centralizada: Implica el mismo lugar pero diferente tiempo. Ejemplo de esto es un foro de debate donde las personas aportan sus comentarios.
- Interacción Síncrona Distribuida: Implica el mismo tiempo pero diferente lugar. Ejemplos de esto son entornos de trabajo, el chat y la video conferencia.

- Interacción Asíncrona Distribuida: Implica diferente tiempo y lugar. Ejemplo de esto es el correo electrónico.

3.5. Tipos de presentación

Se describen tres tipos de situaciones colaborativas que pueden ocurrir en la interfaz de los usuarios.

- *Igual Tarea – Igual Vista*: Esta situación involucra una interacción cercana y requiere awareness del lugar preciso y las acciones exactas de los otros usuarios, La vista igual en los sistemas groupware es llamado "strict WYSIWIS" ('What You See Is What I See')
- *Igual Tarea – Diferente Vista*: Esta situación involucra acciones coordinadas que ocurren en diferentes áreas del espacio de trabajo. La vista diferente en un espacio de trabajo común es llamado "relaxed WYSIWIS", también en este tipo de presentación se incluye el WYSIWID ('What You See Is What I Do').
- *Igual Tarea – Situación de Enfoque Mixto*: Esta situación involucra actividades individuales y compartidas en el espacio de trabajo se entrelazan, y los usuarios desplazan su atención periódicamente entre vistas separadas y compartidas en el espacio de trabajo.

3.6. Métodos de presentación

Son componentes gráficos o herramientas que se muestran en la interfaz del usuario que dan soporte a la diferente información de awareness, estos componentes ofrecen una amplia variedad de información que pueden ser situados dentro o fuera de los objetos compartidos en los sistemas CSCW, son representaciones literales o simbólicas de las acciones de otra persona, estos componente pueden incluir imágenes, video, indicadores de actividad o historial de actividades por mencionar algunas, estos componentes pueden proporcionar una transición entre detalles locales o una vista global y también pueden contener múltiples puntos de foco que pueden ser usados para mostrar detalles de las acciones de cada usuario.

4. Comparación de Mecanismos de Awareness Estudiados

Con base en las características definidas en la sección 3 se ha realizado un estudio comparativo de los distintos mecanismos de awareness relevados cuyo resumen se presenta en la Tabla 1.

Del estudio comparativo realizado se observa que el soporte a los diferentes tipos de awareness exige una combinación de vistas personalizadas y perspectivas sobre diversos objetivos (objetos, personas) en diferentes contextos de uso, dado que la mayoría de mecanismos de awareness relevados tienen un tipo de presentación (igual tarea – situación de enfoque mixto) en donde los usuarios desplazan su atención periódicamente entre vistas separadas y compartidas en el espacio de trabajo, y

presentan la información a través de diversos métodos de presentación (Widget Globales, líneas de tiempo, líneas de conexión, gráficos de estado).

Tabla 1. Comparacion de Mecanismos de Awareness

Mecanismos de Awareness / Características	Nombre del mecanismo	Recolección de datos		Distribución		Presentación		
		Tipos de Awareness	Arquitectura auxiliar	Modo de Awareness	Propagación del mensaje	Tipo de aplicación	Tipo de presentación	Método de presentación
[Gutwin et al., 1995]	Software colaborativo y educativo	Awareness social, Awareness de tareas, Awareness de conceptos, Awareness de espacios de trabajo		Síncrono y Acoplado.	Multi - destino	interacción sincrónica distribuida	1. Igual tarea – igual vista (Strict WYSIWIS) 2. Igual tarea – diferente vista (Relaxed WYSIWIS o WYSIWID) 3. Igual tarea – situación de enfoque mixto	1. Cursos múltiple, cursor semántico 2. Scrollbar multiusuario, widget global 3. mecanismos de historial
[Carroll et al., 2002]	Sistema de notificaciones	Awareness de actividad, Awareness social, Awareness de Acción		Asíncrono y acoplado.	Multi –destino	Interacción asíncrona centralizada	Igual tarea – situación de enfoque mixto	Widget global. Línea de tiempo orientada a actividades hechas por el grupo. Línea de tiempo de actividades sobre un documento
[Collazos et al., 2006]	Sistema multi-agente de enseñanza	Awareness del conocimiento	Agentes inteligentes para la recolección de acciones, envió de reportes y sugerencias	Síncrono y Acoplado. Síncrono y desacoplado. Asíncrono y acoplado. Asíncrono y desacoplado.	Multi –destino	Interacción cara a cara, interacción sincrónica distribuida, Interacción asíncrona centralizada, Interacción asíncrona distribuida	Igual tarea – situación de enfoque mixto	Widget global. Tablero (memoria global). Mail: mensaje directo. sugerencias y reportes.
[Gutwin y Greenberg. 2002]	Groupware de tiempo real	Awareness de espacios de trabajo Awareness de situación			Multi –destino			Lista de participantes, telepointers, avatars, Vistas duplicadas; vistas esclavas
[Collazos et al., 2002]	Mecanismo de conocimiento compartido	Awareness de conocimiento - compartido	Modulo analizador de acciones Modulo analizador de mensajes. Agente generador de Awareness de conocimiento compartido	Síncrono y Acoplado.	Multi-destino		Igual tarea – situación de enfoque mixto	Widget global. icono grafico que puede representar 3 posibles estados: pobre, bueno o alto
[Fuchs et al.,1995]	Mecanismo de eventos locales	Awareness de usuario	Administrador de eventos	Síncrono y Acoplado. Síncrono y desacoplado. Asíncrono y acoplado. Asíncrono y desacoplado.	Relación usuario - objeto	Interacción cara a cara, interacción sincrónica distribuida, Interacción asíncrona centralizada, Interacción asíncrona distribuida	Igual tarea – situación de enfoque mixto	Diferentes colores de iconos. Líneas de conexión con color entre el actor y el objeto. niveles de urgencia del contexto de interes

También se puede observar que la mayoría de mecanismos de awareness comparados realizan la propagación del mensaje a todos los usuarios del sistema (multi-destino), con excepción del mecanismo de eventos locales, el cual presenta una propagación diferente y que puede resultar mas útil para los usuarios, ya que este mecanismo propone una relación usuario-objeto , en donde se define por los contextos de interés, que consisten en un conjunto de tipos de relación, un conjunto de tipos de eventos y una lista de los usuarios interesados que se han suscrito en el contexto. Para cualquier clase de objeto determinado en el sistema, el usuario puede definir (y/o suscribirse a) un contexto de interés.

Otra observación a formular es que los mecanismos de awareness que usan una arquitectura auxiliar dan mayor cobertura a los diferentes modos de awareness (síncrono, asíncrono, acoplado y desacoplado), también implementan todos los tipos de interación proveyendo interacciones cara a cara, asíncronas centralizadas, síncronas y asíncronas distribuidas.

5. Propuesta de Taxonomía de Mecanismos de Awareness

De los mecanismos de awareness relevados y del estudio comparativo realizado, con base en una variación de la propuesta de análisis que se formula en [Gutwin y Greenberg, 2002] se propone un conjunto básico de conceptos que dan respuesta a las preguntas "quién, qué, dónde, cuándo y cómo" para el awareness en espacios de trabajo, con este conjunto básico se identifican elementos específicos de conocimiento que construyen el núcleo de awareness de espacios de trabajo y la lista de las preguntas que cada elemento puede responder. Los elementos son todas las cuestiones de sentido común que tienen que ver con las interacciones entre la persona y el entorno de trabajo. Con base en este conocimiento, se propone una taxonomía de los mecanismos de awareness centrándose en su definición y describiendo que “busca caracterizar” y que “trata de responder” cada tipo. La taxonomía propuesta se resume en la Tabla 2.

Tabla 2.a. Tipos de Awareness

Awareness	Definición	Busca Caracterizar	Tratando de responder a:
social	Información que una persona mantiene sobre otros en un contexto social o de conversación: si una persona está prestando atención, su estado emocional y su nivel de interés.	Intenciones	¿qué debo esperar de otros miembros de este grupo?
		Acciones	¿cómo voy a interactuar con este grupo?
		Habilidades	¿qué rol voy a tomar en este grupo?
			¿qué roles van a tomar los miembros de este grupo?
De Tareas	Es el conocimiento de cómo se completará la tarea	conocimiento	¿qué conozco acerca de este tópico y la estructura de la tarea?
			¿qué conoces los otros acerca de este tópico y la tarea?
		Acciones	¿qué pasos debo tomar para completar la tarea?
			¿cómo el resultado será evaluado?
		Artefacto	¿qué herramientas/materiales son necesarias para completar la tarea?
			¿qué tanto tiempo es necesario? ¿cuanto tiempo esta disponible?

Tabla 2.b. Tipos de Awareness

De Conceptos	De cómo una determinada actividad o conocimiento encaja en el conocimiento existente del estudiante / miembro del grupo	conocimiento	¿cómo esta tarea encaja dentro de lo que ya se, acerca del concepto?
			¿qué mas necesito para encontrar acerca de este tópico?
		Acciones	¿necesito revisar algunas de mis ideas actuales, en base a esta nueva información?
			¿puedo crear una hipótesis de mi conocimiento actual para predecir, la tarea que viene?
De Espacio de Trabajo	Preocupaciones de presencia del usuario en el área de trabajo y lo que los usuarios están haciendo en la actualidad: las interacciones de hasta-al minuto-conocimiento acerca de otras personas con el espacio de trabajo	Acciones	¿qué están haciendo los otros miembros del grupo para completar la tarea?
			¿qué están haciendo?
		Acciones anteriores	¿qué han terminado ellos?
		Ubicación	¿dónde están ellos?
		Intenciones	¿qué es lo siguiente que van hacer?
			¿cómo puedo ayudar a otros miembros de grupo a completar el proyecto?
De Construc-ción del Conoci-miento (individual)	Corresponde a la información que necesitan las personas obtener con el fin de estar al tanto de sus propios conocimientos.	Acciones / Expectativas	¿lo que estoy haciendo esta ayudando a resolver el problema?
		Artefacto	¿necesito mas tiempo/recursos?
		Conocimiento	¿qué mas necesito saber acerca de este tópico?
			¿cuánto tiempo hay disponible?
		Acciones / Expectativas	¿lo que hice esta ayudando a resolver el problema?
		Conocimiento / Acciones / Habilidades	¿qué y como aprendí, de los otros miembros del grupo?
		Conocimiento / Nivel de actividad	¿termine el trabajo?
		Conocimiento	¿qué estoy aprendiendo del grupo de trabajo?
			¿qué necesito saber acerca del tópico?
De Construc-ción del Conoci-miento Compartido (Grupo)	Corresponde a la información que es necesaria con el fin de estar al tanto de los conocimientos de los otros miembros del grupo	Acciones	¿qué están haciendo los otros miembros del grupo, para completar la tarea?
		Acciones/Expectativas	¿lo que los otros están haciendo, esta ayudando a resolver la tarea?
			¿cómo puedo ayudar a otros miembros a completar la tarea?
		Conocimiento	¿qué saben los otros miembros acerca el tópico? ¿qué necesitan saber los otros miembros acerca del tópico?
		Conocimiento / Expectativas	¿qué aprendieron los otros miembros del grupo de mi?
		Ubicación	¿dónde están los otros miembros del grupo?
De Actividad	Hace hincapié en la importancia de los factores contextuales de actividad como la planificación y la coordinación. es el awareness que apoya el desempeño del grupo en tareas complejas. Las actividades son esfuerzos a largo plazo dirigidas a objetivos importantes.	Expectativas	¿Cómo van las cosas?
		Cambios	¿Qué cambios hay en los planes compartidos?
			¿Qué modificaciones hay en los roles del proyecto?
De Acción	Describe avances en las tareas aisladas, proporciona información acerca de las interacciones de otros usuarios con los objetos del espacio de trabajo	Acciones	¿Qué esta pasando?
		Artefactos	¿Qué objetos están usando?
		Cambios	¿Qué cambios están haciendo y donde?
		Nivel de actividad	¿Qué tan activos son en el espacio de trabajo? ¿Qué tan frecuente ocupan un recurso?
De Situación	El conocimiento minuto a minuto requerido para operar o mantener un sistema. la percepción de los elementos relevantes del entorno. Comprensión de esos elementos y predicción del estado de esos elementos	Cambios en el Entorno	¿Qué cambios están ocurriendo en el entorno y donde?

6. Conclusiones

El trabajo mediado por tecnología tiene una evolución de más de dos décadas y ha sido una preocupación de la comunidad académica del área proveer mecanismos de sincronización de la tarea que desarrollan los grupos de trabajo en la que los individuos no comparten el mismo espacio físico para realizarla. Entre estos mecanismos ocupa un rol preponderante el awareness.

Disponer de una taxonomía de los mecanismos de awareness se encuadra dentro de las tendencias actuales de proveer procesos de diseño de espacios virtuales de trabajo personalizables que requieren ajustarse estrictamente a las necesidades de trabajo virtual del grupo [Rodríguez y García-Martínez, 2013].

Como futuras líneas de trabajo se prevé: [a] revisar las características propuestas para describir los mecanismos de awareness y eventualmente agregar nuevas con base en la hipótesis que es necesario enriquecer la descripción de estos mecanismos de manera de permitir la definición de comportamientos correspondientes a situaciones específicas; [b] revisar la taxonomía de mecanismos de awareness propuesta con base en la hipótesis de la posibilidad de su ampliación al identificarse nuevos tipos; y [c] realizar una comparación de las ventajas y desventajas de cada uno de los mecanismos estudiados.

7. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A166 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

8. Referencias

Carroll, J., Neale, D., Isenhour, P., Rosson, M., & McCrickard, S. 2002. *Notification and awareness: synchronizing task-oriented collaborative activity*. Human-Computer Studies, 605–632.

Collazos, C., Builes, J., & Carranza, D. 2006. *Model for supporting awareness in the CSCL ALLEGRO environment through a blackboard architecture*. Revista de Ingeniería e Investigación , pp. 67-77.

Collazos, C., Guerrero, L., Pino, J., Ochoa, S. 2002. Introducing Shared-Knowledge Awareness. International Conference: Information and Knowledge Sharing, pp. 13-18.

Cuyas, A. 1982. *Appleton-Cuyas Spanish English/English Spanish Dictionary*. Prentice Hall. ISBN 10: 0136155596

Dourish, P.; Bellotti, V. 1992. *Awareness and Coordination in Shared Workspaces*. In Proceedings of the 1992 Conference on Computer-Supported Cooperative Work, 107-114.

Ellis, C., Gibbs, S., Rein, G. 1991. Groupware: Some Issues and Experiences. Communications of the ACM, Vol. 34 No. 1.

Fuchs, L., Pankoke-Babatz, U., Prinz, W. 1995. *Supporting Cooperative Awareness with Local Event Mechanisms: The GroupDesk System*. Proceedings of the Fourth European Conference on Computer–Supported Cooperative Work, ECSCW'95 , pp. 247–262.

Gutwin, C., Stark, G., & Greenberg, S. 1995. *Support for Workspace Awareness in Educational Groupware*. Proceedings Conference on CSCL, pp. 147-156. LEA Press.

Gutwin, C.; Greenberg, S. 2002. *A Desriptive Framework of Workspace Awareness for Real-Time Groupware*. Computer Supported Cooperative Work: The Journal of Collaborative Computing , 411-446.

Harper Collins, 2011. *Concise English-Spanish Dictionary*. Harper Collins Publishers. ISBN 10: 0061141844.

Rodríguez, D., García-Martínez, R. 2013. *Propuesta de Proceso de Diseño de Espacios Virtuales de Trabajo Educativo Personalizables*. Proceedings VIII Congreso de Tecnología en Educación y Educación en Tecnología. ISBN 978-987-1676-04-0.

WebECALEAD: diseño de un prototipo web como herramienta de soporte para la Evaluación de Calidad en Educación a Distancia

Gladys Gorga[1] , Cecilia Sanz[1], Cristina Madoz[1]

[1] Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata. 50 y 120, La Plata, Buenos Aires, Argentina

{ggorga, csanz, cmadoz}@lidi.info.unlp.edu.ar

Abstract. Se presenta la descripción del diseño de un prototipo de sistema web para la evaluación de calidad en Educación a Distancia, denominado WebECALEAD, que tiene como objetivo facilitar la tarea de evaluación de los procesos más destacados que intervienen en los modelos educativos a distancia o híbridos, basado en la propuesta ECALEAD desarrollada por lo autores. El prototipo contempla las fortalezas y debilidades de ECALEAD, en base a su aplicación en experiencias educativas concreta. Se presenta aquí una introducción a la temática y algunos antecedentes, la descripción del prototipo diseñado, un análisis de las posibilidades que ofrece éste en función del modelo ECALEAD. Finalmente, se presentan las conclusiones y trabajos futuros.

Keywords: calidad, educación, modelo de evaluación, sistemas web para evaluación de procesos

1 Introducción

Los actuales escenarios educativos, están fuertemente impregnados de la revolución tecnológica, y proponen un verdadero desafío para las instituciones de educación superior, que deben aprovechar el potencial de los recursos tecnológicos para ofrecer propuestas educativas de calidad. En este sentido, las instituciones destinan, parte de sus esfuerzos, al diseño e implementación de nuevas formas de organización y gestión de sus ofertas, incorporando innovación tecnológica para hacer frente a los nuevos desafíos. Surge entonces la necesidad de establecer mecanismos para asegurar la calidad, tanto de la oferta pública como privada en el ámbito de la Educación Superior. En varios países de América Latina se vienen implementando algunas políticas en este sentido. Así es que las instituciones de Educación Superior (IES) se han comprometido a realizar procesos de autoevaluación, someterse a procesos de acreditación y emplear indicadores de desempeño, entre otras acciones [1][2][3][4].

En este sentido, resulta necesario que las IES fijen claramente los criterios y estándares de calidad a seguir para alcanzar sus fines.

Para determinar el conjunto de criterios de calidad de las IES, en cualquiera de las modalidades presencial o a distancia, es fundamental analizar cuál es el contexto

particular en el que se desarrollan, cuáles son sus componentes principales, los aspectos críticos, los actores que formarán parte, y sus características, necesidades y demandas, entre otros.

En [5] se presentó ECALEAD, un modelo para evaluar la calidad de un sistema / programa/ proceso educativo a distancia, que integra algunos de esos aspectos poniendo el foco específicamente en los procesos más importante involucrados a criterio de sus autores. Este modelo puede ser utilizado también en modalidades educativas híbridas que combinan aspectos de la modalidad presencial con estrategias de trabajo y aprendizaje a distancia.

En [6] [7] se presentó un análisis de las bondades y debilidades que ofrece el modelo ECALEAD derivadas de su aplicación en algunas experiencias educativas concretas. Este trabajo ha ido evolucionando, y actualmente se está trabajando en el sistema foco de este trabajo.

A partir de estas consideraciones se presenta ahora la descripción del diseño de un prototipo de sistema web denominado WebECALEAD, y tiene como objetivo facilitar y acompañar el proceso de evaluación propuesta por ECALEAD.

2 Breve descripción de los aspectos de evaluación de ECALEAD

El modelo ECALEAD, sigue los lineamientos planteados por García Aretio, y propone un modelo integrador, de desarrollo y control de la calidad total, vinculado al contexto, metas, entradas, procesos, resultados, y mejoras. En ese contexto se establecen algunos criterios generales como funcionalidad, disponibilidad, eficacia, eficiencia, información e innovación [8].

ECALEAD es un modelo de evaluación de calidad basado en tres capas que profundizan en la tarea:

En la **Capa 1** se definen los procesos a evaluar y los criterios generales a considerar en el modelo. **S**e consideran los criterios mencionados, y se los relaciona con las metas/objetivos, procesos, recursos, resultados y mejoras involucrados en el objeto de evaluación, que se muestran en la Fig. 1.

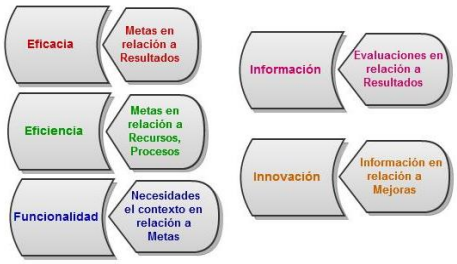


Fig. 1. Criterios a considerar en la capa 1, en función de los componentes vinculados a una IES, en el marco de este trabajo

En la **Capa 2** se proponen algunos indicadores que se relacionan en forma directa con los criterios y procesos de la primera capa. Para ello, deben determinarse los recursos y resultados, que se consideren de interés en un sistema/programa/curso a

distancia. Al aplicar la capa 2 en una evaluación específica, se deben ajustar estos indicadores propuestos por el modelo, acorde al contexto. No siempre será necesario aplicar todos los indicadores propuestos, o tal vez, en algunos casos se requiera involucrar nuevos indicadores. El modelo aborda un conjunto de indicadores que se creen pertinentes en una evaluación de calidad de sistemas de EAD, en general.

A modo de ejemplo se detallan algunos de los indicadores correspondientes a uno de los procesos que ECALEAD propone, vinculándolos con los criterios generales de la Capa 1. El lector puede observar la descripción completa de procesos, criterios e indicadores en [9]

El ejemplo que se presenta aquí es el del proceso de **Administración y Gestión** que incluye, entre otros aspectos, establecer circuitos de difusión, inscripción, atención de consultas administrativas, gestión de alumnos y docentes, tales como mantener, registrar y dar información sobre cursos aprobados, notas, certificaciones alcanzadas, etc. También se involucran los circuitos para entregar credenciales de acceso a entornos virtuales de enseñanza y aprendizaje (en caso de haberlos). Algunos Indicadores a considerar para este proceso son

Vinculados a los criterios de eficacia y eficiencia:

Cantidad de inscripciones (analizando procedencia del alumno) que se reciben, en relación con la cantidad de recursos humanos que atienden dichas inscripciones y circuitos disponibles para la inscripción presencial o a distancia.

Cantidad de alumnos que inician el proceso luego de la inscripción, en relación con la cantidad de inscriptos. Pueden indicar faltas de información administrativa adecuadas.

Cantidad de trámites administrativos iniciados en el día en relación con los finalizados.

Porcentaje de consultas recibidas a través de los diferentes medios disponibles.

Vinculados al criterio de funcionalidad:

Cantidad de alumnos y docentes que manifiestan satisfacción respecto de los circuitos administrativos y de consulta disponibles en la institución.

Vinculados al criterio de información e innovación:

Verificación de que existe un informe que reporte los resultados obtenidos a partir de los indicadores.

Cantidad de mejoras y cambios implementados en vinculación con el plan de mejoras determinado por la institución.

Los indicadores mencionados constituyen un conjunto de ejemplo, que no es exhaustivo, pero permite orientar el tipo de trabajo a realizar en la capa 2. Como ya se explicó, la definición concreta de indicadores es establecida acorde al objeto de evaluación y su contexto [10].

En la **Capa 3** se propone tomar cada uno de los indicadores para los procesos y criterios en cuestión, y analizarlos de manera tal, de concretar la evaluación para el objeto particular a considerar (curso, sistema, proyecto de EAD). Esta capa es la más específica y debe ajustarse al contexto particular. Como podrá observarse en el desarrollo que se ha realizado en la capa 2, existen algunos indicadores que son cuantitativos mientras que hay otros más cualitativos. La definición de escalas para la determinación de calidad, dependerá de esta característica, en cada caso. Es en esta capa donde se determinan las medidas a considerar con sus escalas específicas.

3 Análisis de fortalezas y debilidades del modelo propuesto

A partir de la aplicación de ECALEAD a experiencias educativas concretas, se ha observado que la flexibilidad del modelo obliga a revisar detalladamente con qué procesos, criterios e indicadores se abordará la evaluación para que puedan ser accesibles y/o adecuados acorde al objeto de estudio particular. Por otra parte, si la evaluación es realizada por personal más vinculado a los niveles organizativos o de gestión de una institución, se contemplará la posibilidad de considerar mayor cantidad de procesos en detalle, como el de políticas y normativas, el vinculado con lo económico y financiero, entre otros. Sin embargo, si los evaluadores son docentes de un curso particular, es probable que no puedan considerarse políticas y normativas propias del curso sino las propias de la institución que lo gestione/organice. En esos casos, es probable que el docente no tenga acceso a información para llegar a analizar dichos procesos, y sólo se trabajará con los más relacionados con el diseño del curso en sí. Es por ello, que la aplicación del modelo deberá ajustarse acorde a las posibilidades de cada contexto específico, y de quiénes llevarán a cabo el proceso de evaluación.

De acuerdo a lo expresado en [7], se detallan algunas de las fortalezas y debilidades encontradas en ECALEAD y que sirven como punto de partida para el presente trabajo:

Fortalezas

- Flexibilidad para adecuar el modelo acorde a las necesidades del contexto y objeto a evaluar
- Gradualidad en las decisiones. Esto permite dividir la tarea, ordenarla y concentrarse en cada momento en los aspectos propios de cada capa.
- Especificidad en procesos educativos mediados por tecnología digital. El modelo tiene en cuenta aspectos directamente relacionados con el ámbito educativo, y particularmente, aborda el análisis de modalidades educativas híbridas o a distancia.
- Pertinencia para la evaluación de calidad. La aplicación del modelo completo permite dar cuenta de una variedad de cuestiones que hacen a la calidad del objeto evaluado.

Debilidades

-Variedad de indicadores (cualitativos y cuantitativos), lo cual puede dificultar la definición de medidas para determinar calidad, si el evaluador desconoce cómo medir cada uno.
- Falta de definición del modelo respecto de qué medidas es conveniente adoptar en cada caso. Sólo se presentan un conjunto de indicadores para cada criterio, pero este conjunto puede no abarcar algunas situaciones importantes para el objeto a evaluar, que quedan en manos de quienes lleven adelante la evaluación. Esto puede provocar, omisiones importantes en la evaluación, si no se lo ajusta adecuadamente.
- Cantidad de recursos humanos involucrados. Llevar adelante el modelo completo, implica un trabajo que involucra a diferentes recursos humanos. Desde quienes abordan las decisiones para llevar adelante esta evaluación, hasta involucrar a los que aportan información para realizarlo, de manera adecuada y lo más completa posible. La calidad de los instrumentos utilizados para recoger información, influirán notablemente en los resultados de la evaluación.

Este análisis lleva a los autores a elaborar una propuesta superadora que propone un prototipo de un sistema web que permitirá desarrollar el proceso de evaluación propuesto por ECALEAD, considerando las fortalezas y debilidades observadas a partir de la aplicación del modelo a las diferentes experiencias. De esta manera, se propone acompañar el proceso propuesto por ECALEAD a partir de un sistema que ayude en la toma de decisiones, establecimiento de medidas, entre otros aspectos.

4 Características del prototipo

Se presenta aquí un prototipo de un sistema web cuyo objetivo principal es asistir al evaluador en el análisis de calidad de procesos educativos mediados por tecnología. Esto conlleva la definición de criterios, indicadores y medidas para cada uno de los procesos contemplados en ECALEAD, conforme a lo detallado en cada una de las capas de este modelo de evaluación. Este trabajo gradual que avanza a través de las capas, de lo general a lo específico, estará directamente afectado por el rol y las responsabilidades del evaluador como integrante de la institución educativa.

Se ha decidido que sea web, dado que esto permite aprovechar ventajas tales como la portabilidad, que facilita su ejecución desde cualquier computadora con conexión a internet y su funcionalidad independiente del sistema operativo instalado en el equipo del evaluador, entre otras.

Por otra parte, se contemplan dos perfiles de trabajo: un perfil de usuario administrador y otro de usuario evaluador. El usuario administrador se encargará de definir los objetos de evaluación, los roles de evaluadores, los procesos, criterios e indicadores que contempla ECALEAD. Este perfil no se aborda en la presentación del trabajo. En cuánto al usuario evaluador, el prototipo permite, de acuerdo al rol asignado (docente, directivo, etc.), armar un plan de evaluación, a través de un esquema que se genera gradualmente, al que van agregando aquellos procesos, criterios, indicadores y medidas que el evaluador considere de interés, acorde al contexto y la intención de la evaluación.

El diseño del prototipo presenta dos áreas principales para organizar el plan de evaluación:

- **Área Superior** que está orientada a asistir al evaluador respecto de la capa de evaluación en la que trabaja y las tareas involucradas. Está compuesta por solapas que al mismo tiempo sirven como elementos de navegación.
- **Área Inferior** donde se presentan los elementos que el evaluador va seleccionando a lo largo del proceso. En todo momento podrá ver el avance del plan de evaluación a través del esquema generado en esta área.

A continuación, se presentan sintéticamente las distintas etapas que el prototipo propone para el armado del plan de evaluación.

4.1 Plan de evaluación en el prototipo

El plan de evaluación se organiza de modo que el evaluador avanza a través de las distintas capas que propone ECALEAD, asistiéndolo en la recolección de los datos

para una experiencia concreta, y brindando la posibilidad de observar los resultados finales de la evaluación. Los pasos a seguir pueden sintetizarse como:

Registrarse: el evaluador debe registrarse a través de un formulario y la institución le asigna un rol específico. El sistema admite los siguientes roles: evaluador docente, evaluador administrativo y/o evaluador directivo (Fig. 2). De acuerdo a este rol, se le habilita la evaluación de determinados procesos, criterios e indicadores.

Fig. 2 Ingresa los datos de registración en el sistema

Seleccionar actividad: el evaluador debe seleccionar la actividad que llevará a cabo y el objeto de la evaluación. Las opciones que se presentan son: a- Iniciar Evaluación, b- Continuar Evaluación, c- Ver Resultados Evaluación (Fig. 3).
Tanto en la opción de **Iniciar Evaluación** como en la de **Continuar Evaluación** se dispone de información sobre la capa de ECALEAD en la que se está trabajando, y las tareas involucradas en dicha capa.
En **Ver Resultados Evaluación** se muestra el plan de evaluación completo, distinguiendo la situación particular de cada indicador, es decir, si ha alcanzado o no la medida de calidad esperada.

Fig.3 Selecciona la actividad a realizar en el sistema

Crear Evaluación: el armado del plan de evaluación comienza por la Capa 1 según ECALEAD. Se selecciona el objeto de la evaluación (curso, programa, carrera, etc). El evaluador debe seleccionar los procesos a analizar. Para cada proceso establece la relación con el conjunto de criterios que analizará. Las relaciones elegidas se agregan a la tabla de procesos y criterios. Cada entrada en la tabla permite acceder al conjunto de indicadores correspondientes (Fig.4). Es objetivo del sistema, que en un trabajo

futuro permita la incorporación de otros criterios y procesos no contemplados en ECALEAD, al momento.

Fig. 4 Arma la tabla de procesos y criterios a analizar

Seleccionar los indicadores: en la Capa 2, el evaluador debe seleccionar los indicadores para cada entrada de la tabla de procesos y criterios generada en la Capa 1. El sistema permite incorporar nuevos indicadores que no aparecen descriptos y que se consideren pertinentes. En esta etapa, el evaluador puede optar por una de las siguientes acciones que siempre estarán presentes como elementos de navegación:

- **Continuar** que permite avanzar a la siguiente etapa (en este caso a la realización de tareas correspondientes a la Capa 3)
- **Anterior** que permite ir a la etapa anterior. En este caso sería revisar las entradas de la tabla de procesos y criterios. Esto puede conducir a otra selección de Indicadores.
- **Salir** que permite guardar el trabajo realizado hasta el momento y retomarlo posteriormente, a través de la opción Ver Resultados Evaluación (Fig. 5).

En todos los casos, el evaluador también utiliza el área superior como herramienta de navegación para ir a una determinada etapa.

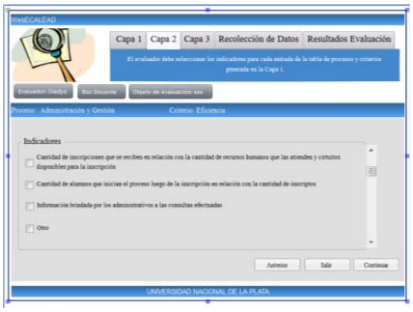


Fig.5 Selecciona los indicadores apropiados, a criterio del evaluador

Seleccionar las medidas: en la Capa 3 se presentan los indicadores elegidos, y se deben establecer las medidas más adecuadas al contexto y objeto de evaluación. En los casos de indicadores cuantitativos, se permite el ingreso de valores correspondientes a la cota inferior y superior del intervalo. En otros casos, se deberán utilizar escalas cualitativas ya estandarizadas (disponibles en WebECALEAD) o emplear la opción de incorporar una nueva escala. Luego, el evaluador puede optar por ir a la etapa de Recolección de Datos (Fig. 6) que se explica a continuación.

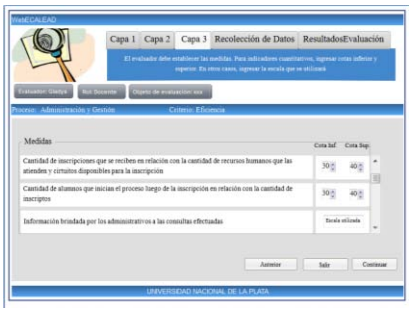


Fig.6 Ingresa las medidas acordes a cada indicador seleccionado

Recolectar Datos: en esta etapa, se ingresan los valores para cada medida, obtenidos a partir de los diferentes instrumentos de recogida de datos que se hayan utilizado. En esta etapa, los valores debieran coincidir con aquellos de la escala seleccionada. Como último paso puede optar por visualizar el plan de evaluación completo con el estado de cada proceso o continuar con el armado del plan (Fig. 7).

Fig.7 Ingresa los valores obtenidos para cada indicador analizado

Ver Resultados Evaluación: aquí se observa el plan de evaluación completo. Se da la posibilidad de elegir un determinado proceso de los que componen el plan y uno de sus criterios. Esta acción despliega la lista de indicadores asignados a ese criterio, y se muestra el estado de cada uno a partir del uso de colores de referencia: verde cuando el indicador obtuvo un valor que cae dentro del rango de valores esperados, amarillo si el indicador no ha sido evaluado aún o rojo si el indicador se evaluó completamente, pero no alcanzó los resultados esperados (Fig. 8).

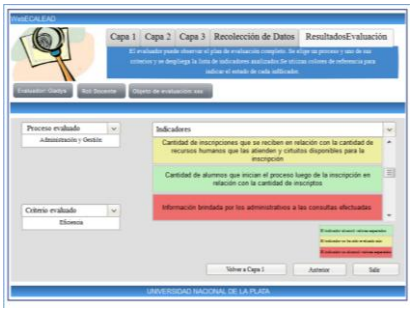


Fig. 8 Visualiza los resultados de la evaluación para un determinado proceso y criterio

4.2 Análisis de beneficios del prototipo

En esta sección se presentan los resultados de un primer análisis realizado por los autores, en relación a las posibilidades que ofrece el prototipo, tomando en cuenta el estudio previo de fortalezas y debilidades, ya realizado para ECALEAD. En trabajos futuros se someterá el prototipo a evaluación por medio de juicio de expertos.

Respecto de las fortalezas de ECALEAD el prototipo las considera, y respeta, dado que:

- adecua el modelo a las necesidades del contexto y objeto a evaluar. Avanza gradualmente en la toma de decisiones partiendo de los procesos y llegando a los valores concretos para cada indicador, lo que permite dividir la tarea, ordenarla y concentrarse en cada momento en los aspectos propios de cada capa.
- los procesos, criterios e indicadores contemplados posibilitan el análisis de modalidades educativas híbridas o totalmente a distancia, a diferencia de otros modelos que son aplicables a la modalidad presencial, y por lo tanto, dejan de lado cuestiones propias de estas modalidades.
- la aplicación del modelo completo puede dar cuenta de todos los aspectos que hacen a la calidad del objeto evaluado.

Por otra parte, el diseño del prototipo se focaliza en corregir algunas de las debilidades de ECALEAD, logrando:

- definir el conjunto de medidas convenientes a adoptar en cada caso
- adaptar el prototipo a la variedad de indicadores (cualitativos y cuantitativos), facilitando la definición de medidas adecuadas a cada uno
- permitir la incorporación de nuevos indicadores no contemplados en las listas sugeridas, de manera tal de incluir situaciones no especificadas por ECALEAD.

Al mismo tiempo se han encontrado funcionalidades beneficiosas que aporta el prototipo tales como: la visualización del plan completo de evaluación, y su estado de avance en la obtención de resultados, la posibilidad de integrar instrumentos para la evaluación a través de los cuales se importen directamente los valores para cada indicador, y la sistematización general del proceso de evaluación propuesto por ECALEAD que ofrecerá un resguardo de la información pertinente.

5 Conclusiones y trabajos futuros

Se ha presentado el diseño de un prototipo web basado en el modelo ECALEAD creado por los autores de este trabajo, que tiene como objetivo facilitar la tarea de evaluación de los aspectos más destacados que intervienen en los procesos educativos que incluyen el uso de tecnología digital.

El diseño presentado respeta las fortalezas de ECALEAD y corrige algunas de sus debilidades, permitiendo completar el plan de evaluación acorde al rol y responsabilidades del evaluador. Al mismo tiempo agrega las ventajas propias de un sistema informático web que se vinculan con las posibilidades de acceso y portabilidad, entre otras.

En cada una de las capas propuestas por ECALEAD, el evaluador puede revisar y

corregir algunas decisiones tomadas en etapas anteriores, actualizando los aspectos que se trabajan en ellas, es decir el proceso en el plan de evaluación es dinámico y facilita la toma de decisiones.

La evaluación final muestra el plan de evaluación completo distinguiendo la situación particular de cada proceso con sus correspondientes criterios e indicadores analizados, de modo de revisar rápidamente cuales son los aspectos débiles encontrados en la evaluación, cuáles faltan completar y cuáles son sus fortalezas en relación al objeto de evaluación.

Como trabajo a futuro se propone:

- la evaluación del diseño del prototipo presentado por medio de la técnica de juicio de expertos de manera tal de evolucionar el prototipo a un sistema final
- la implementación del sistema web considerando decisiones de diseño para respetar criterios de usabilidad
- la incorporación a WebECALEAD de otros procesos y/o criterios no contemplados en el prototipo, y que pueden ser de utilidad para un evaluador particular. Esto implicaría la incorporación de nuevos indicadores que se ajusten a tales procesos y/o criterios
- la creación de instrumentos tales como encuestas, entrevistas, etc. que asistan al evaluador en la recolección de indicadores cualitativos, permitiendo de esta manera que los resultados obtenidos puedan incluirse en la propia base de datos del sistema.

Estos son los primeros pasos en relación al diseño e implementación de WebCALEAD, se espera en el corto plazo tener los primeros resultados de su evaluación.

Referencias

1.Brunner J.J. Nuevas demandas y sus consecuencias para la Educación Superior en América Latina. Santiago de Chile. 2002.

2.Fernandez Lamarra, N. (Compilador). Universidad, Sociedad e Innovación [Una perspectiva internacional] Compilado. Univ. Nacional de Tres de Febrero. 2009.

3.Mena M. (Compiladora) Construyendo la nueva agenda de la Educación a Distancia. Ed. La Crujía. 2007.

4.Brunner JJ, Ferrara Hurtado R. (Coordinadores) Educación Superior en Iberoamérica – Informe 2011.

5.Gorga, Madoz, Sanz. Evaluación de calidad en sistemas de educación a distancia. Propuesta y aplicación a un caso de estudio. III-LIDI Fac. de Informática. UNLP. CACIC 2010.

6.Gorga, Sanz, Madoz, Manresa, Abásolo. ECALEAD: Evaluación de Calidad en Educación a Distancia. Aplicación en un caso de estudio. CAFVIR 2012. España. 2012

7.Gorga, Sanz, Madoz. ECALEAD – Evaluación de Calidad en Educación a Distancia. Análisis del modelo propuesto. III-LIDI. Facultad de Informática. UNLP. CACIC 2011.

8.García Aretio, L. (coord.) et al. "De la Educación a Distancia a la Educación Virtual". Edit. Ariel. ISBN: 978-84-344-2666-5. 2007.

9.Sanz, Gorga, Madoz. Propuesta de un modelo de evaluación en capas. CACIC 2007. Argentina. ISBN: 978-950-656-109-3

10.Sanz, Gorga, Madoz. El tema de la calidad en la educación a distancia. Propuesta de un modelo de evaluación en capas. CACIC 2008. Argentina. ISBN 978-950-9474-49-9

Juegos Educativos Móviles: Aspectos involucrados

Alejandra B. Lliteras[1], Cecilia Challiol[1,2] and Silvia E. Gordillo[1,3]

[1] UNLP, Facultad de Informática, LIFIA, La Plata, Bs. As., Argentina
{lliteras,ceciliac,gordillo}@lifia.info.unlp.edu.ar

[2] Also CONICET, [3] Also CICPBA

Abstract. Los Juegos Educativos Móviles son creados con objetivos educativos y se emplean como una herramienta en el proceso de aprendizaje por considerarse un elemento motivador para los alumnos. En este trabajo se presentan tres aspectos de este tipo de juegos: el contenido, la movilidad y la presentación (que se le brinda al alumno de los dos aspectos anteriores). El objetivo de este trabajo, es presentar estos aspectos de manera desacoplada entre sí, para poderlos combinar, al momento de la creación de estos juegos, fomentando de este modo, el reuso de cada aspecto.

Keywords: Juegos Educativos Móviles, Aprendizaje Móvil, Aplicaciones Móviles Educativas

1 Introducción

Desde hace más de veinte años, se ha planteado en [1] que la forma en la que las instituciones educativas fomentan o imparten el conocimiento, no se condice con la manera en la que los alumnos aprenden fuera de la misma. Se destaca además, que la mayoría de las actividades mentales que realiza un alumno fuera de la institución escolar, a diferencia de las que se realizan dentro de la misma, implican el uso de herramientas, las cuales inciden en su actividad cognitiva resultante. [2] y [3] abordan esta idea cuando, en particular, dichas herramientas son tecnológicas (por ejemplo: los dispositivos móviles, los cuales son usados desde temprana edad por los alumnos [4]).

El uso de dispositivos móviles, como se menciona en [5], ha cambiado la forma en la que se aprende, y la manera en la que los distintos actores involucrados en el proceso de aprendizaje se relacionan entre sí. Este recurso es empleado por los docentes, como una herramienta adicional en dicho proceso [6]. Para esto, los docentes, incorporan el uso, por parte de los alumnos, de aplicaciones móviles educativas. Distintos autores (por ejemplo, [7], [8] y [9])), asumen que los alumnos se encuentran en permanente movimiento mientras reciben, en distintas posiciones (o ubicaciones), contenido educativo en su dispositivo móvil. La movilidad del alumno puede permitirle que, en cada posición, tenga una experiencia de aprendizaje distinta.

En [10], se propone pensar a las aplicaciones de software en el ámbito educativo, como juegos, a fin de motivar a los alumnos en el proceso de aprendizaje. Esto es sustentado por los efectos motivadores de los juegos en general [11]. A los juegos creados específicamente con fines educativos se los conoce como, *"Serious Games"* (Juegos Serios) [12], en particular, cuando éstos se presentan incorporando el aspecto de movilidad, se denominan *"Mobile Serious Games"* (Juegos Serios Móviles) [13] o Juegos Educativos Móviles [14].

En la actualidad, la mayoría de los Juegos Educativos Móviles que contemplan el movimiento del alumno (también conocidos como, basados en posicionamiento), como por ejemplo, Savannah [15], Frequency 1550 [16] y MobileMath [17], son creados de manera ad-hoc, no se puede reusar el contenido, el aspecto de movilidad ni la presentación, es decir, existe acoplamiento entre los aspectos mencionados.

Supongamos un Juego Educativo Móvil basado en posicionamiento (JEMBP), que se juega dentro de un zoológico. Dentro de ese zoológico, un lugar destacado podría ser por ejemplo, la jaula de los leones. Cuando un alumno llega a dicho lugar, recibe en su dispositivo móvil un contenido educativo y un mapa esquemático del área del zoológico donde se desarrolla el juego. La flexibilidad deseada para la creación de este juego, podría involucrar, por ejemplo, reusar: el contenido con distintas presentaciones (por ejemplo, en texto o video, acorde a las capacidades sensoriales del alumno), el mapa esquemático del zoológico para reusarlo en otros juegos, o bien para presentarlo de diferentes maneras (por ejemplo acorde a las capacidades cognitivas de los alumnos que lo reciben o bien, para poder considerar la asistencia al alumno mientras éste se mueve). También podría involucrar, reusar, en el caso de ser posible, el contenido en diferentes zoológicos, como así también, el lugar destacado (la jaula de los leones) para brindar otros contenidos en otros juegos. Por lo ya argumentado anteriormente, esto no es posible en los juegos creados ad-hoc ya que los aspectos involucrados en los mismos, son generados de manera acoplada.

El objetivo de este trabajo, es presentar los aspectos de contenido, movilidad y presentación de manera desacoplada entre sí, para poder combinarlos, al momento de la creación de estos juegos, fomentando de este modo, el reuso de cada aspecto.

Este trabajo se estructura de la siguiente manera: en la Sección 2 se aborda el estado del arte tanto de aprendizaje móvil, como de los Juegos Educativos Móviles. En la Sección 3 se brinda una caracterización de los aspectos involucrados en los Juegos Educativos Móviles de una manera desacoplada entre sí. En la Sección 4 se presentan distintos Trabajos Relacionados y finalmente en la Sección 5, se presentan las Conclusiones y los Trabajos Futuros.

2 Estado del Arte

En esta sección se abordarán dos áreas relevantes para el trabajo propuesto, como son el Aprendizaje Móvil y los Juegos Educativos Móviles, destacando para cada una, distintos trabajos que sirven como sustento teórico.

2.1 Aprendizaje Móvil

El término "aprendizaje móvil" (también conocido por su acrónimo en inglés *mlearnig* o *m-learning*) ha sido abordado, a lo largo del tiempo, desde diferentes perspectivas. Por ejemplo, [18] destaca tres perspectivas: desde lo tecnológico [19], como una extensión del aprendizaje electrónico (conocido por su acrónimo en inglés como *elearning* o *e-learning*) [20] y desde la movilidad [21].

De acuerdo a [22], en el proceso de aprendizaje móvil, el aspecto más importante, no es el tecnológico, sino el de movilidad. En particular, se destaca la importancia de la movilidad en el espacio, ya que el alumno va encontrando oportunidades de aprendizaje, en el ambiente o espacio, a medida que se desplaza en él. En nuestro trabajo, se adoptará la perspectiva desde la movilidad del alumno durante la experiencia de aprendizaje móvil. En particular, se aborda la relevancia en el contenido que un alumno recibe en una aplicación móvil educativa (o juego educativo móvil), acorde a su cambio de posición en el espacio en el que se mueve.

El aprendizaje móvil, ha cobrado relevancia en los últimos años, debido a la popularidad de los dispositivos móviles [8], mediante los cuales se puede hacer llegar contenido educativo a lugares en los que, de otra manera, seria costoso y difícil [23]. El término en cuestión, es aún, de acuerdo a [4], objeto de debate.

2.2 Juegos Educativos Móviles

Como se mencionó en la introducción, en [10] se propone pensar a las aplicaciones de software en el ámbito educativo, como juegos, a fin de motivar a los alumnos en el proceso de aprendizaje. Donde de acuerdo a este autor, existen seis elementos estructurales que caracterizan a un juego: las reglas que organizan el juego, las metas y objetivos, el resultado (*outcome*) y la retroalimentación (*feedback*) que permiten medir el progreso ante las metas. También están, el desafío, la interacción social y la historia.

En este trabajo, se abordarán los JEMBP, donde de acuerdo a [14], el alumno se encuentra en movimiento durante la experiencia del juego. En la Figura 1, se puede apreciar, la relación de inclusión de este tipo de juegos en relación con otros conceptos previamente presentados.

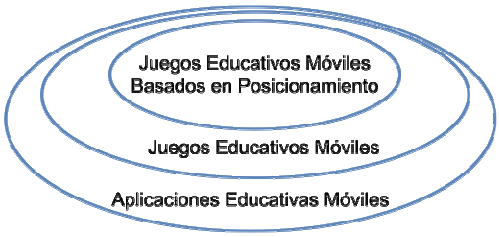


Fig. 1. Juegos Educativos Móviles basados en posicionamiento

De acuerdo al trabajo presentado en [24], los JEMBP encuadran en el patrón de diseño de juegos móviles llamado "Navegación Física". En este tipo de juegos, el

jugador debe moverse en el mundo físico para jugar. En [25] se realiza un análisis de la aplicación de este patrón de juego en relación al resultado de aprendizaje en el participante y la manera en el que el juego impacta en la motivación del mismo. En dicho trabajo se menciona que: el participante está altamente motivado, interesado y en movimiento.

En [26] se presenta una guía para el diseño de JEMBP brindando pautas a tener en cuenta en el momento de la creación y uso de los mismos, sin embargo, no se presenta una clara separación de los aspectos abordados en nuestro trabajo ya que no hacen hincapié en el reuso.

3 Caracterización de los Aspectos involucrados en los JEMBP

En [27] presentamos una guía para conceptualización de los JEMBP, haciendo hincapié en desacoplar los aspectos de contenido y movilidad, en particular, para asistir a un equipo multidisciplinario (por ejemplo, formado por expertos en educación y expertos en tecnología) en la creación de los mismos y fomentar el reuso de estos dos aspectos. En la Sección 3.1 se presentarán las características principales de los aspectos de contenido y movilidad. En la Sección 3.2 incorporamos la noción de presentación como un nuevo aspecto a tener en cuenta en los JEMBP.

3.1 Aspectos de Contenido y Movilidad

Como se menciona en [27] se pueden identificar tres formas de estructurar el contenido en un JEMBP: secuencial lineal, secuencial con bifurcación y conjunto. Para reducir la complejidad en la creación y posterior reuso en este tipo de aplicaciones, se separa en dos capas los aspectos de contenido y movilidad (la cual involucra el espacio de juego y las posiciones destacadas dentro del mismo). Esta separación, permite mejorar la reusabilidad y simplificar las problemáticas que vienen relacionadas con la evolución de cada aspecto antes mencionado.

A continuación, veamos cómo esta separación en capas se puede aplicar al JEMBP del zoológico propuesto previamente. En la Figura 2, se puede apreciar a izquierda un espacio de juego que representa un determinado zoológico, donde se definen posiciones destacadas, en las que se le brindará contenido educativo al alumno en el marco de un juego, en particular, hay dos estructuras de contenidos definidas que se relacionan con las posiciones destacadas para formar así dos JEMBP diferentes, es decir, se está reusando el aspecto de movilidad. Acorde al juego en el que este participando el alumno, será el contenido que reciba en su dispositivo móvil. En la Figura 2, a derecha se puede observar, otro espacio de juego (en este caso otro zoológico), con sus propias posiciones destacadas y reusando el contenido previamente definido (en un JEMBP de otro zoológico), se crea un nuevo JEMBP haciendo reuso del aspecto de contenido pre existente. Este reuso es posible, siempre y cuando los contenidos no hayan sido creados ad-hoc para un lugar destacado específico. Por ejemplo, una pregunta que hace referencia a una característica propia e irrepetible de una jaula de un zoológico en particular.

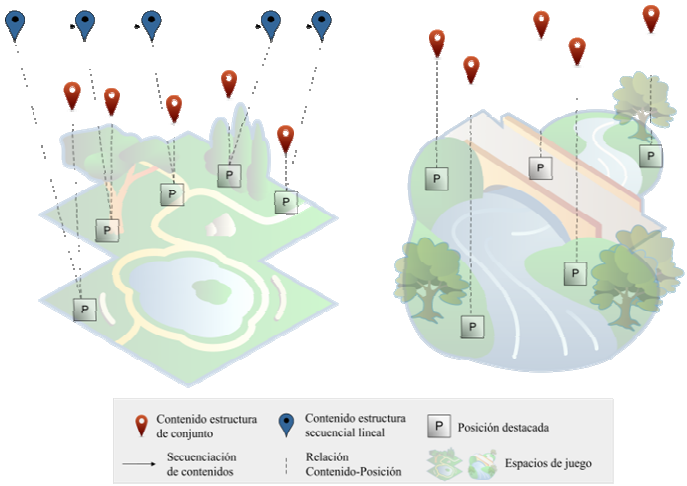


Fig. 2. Reuso de los aspectos de contenido y movilidad

En los aspectos presentados en esta sección no se abordó la temática de cómo presentar los mismos al alumno. Esto será presentado en la siguiente sección.

3.2 Presentación

En este trabajo se identifican dos presentaciones a contemplar en los JEMBP: por un lado la presentación del contenido educativo y por otro la presentación del espacio de juego. Cada una de estas temáticas ha sido abordada por diferentes autores, pero ninguno de ellos las ha integrado de manera conjunta con el fin de generar JEMBP, y tampoco han contemplado la generación de estas presentaciones, para su reuso.

A continuación se presentan trabajos que abordan, por un lado, la presentación de contenido, y por otro, la presentación del espacio de juego.

- El contenido, de acuerdo a [28] puede brindarse en forma de audio, imagen, video y/o texto. La adopción de una de estas formas o de una combinación de ellas, dependerá del grupo de alumnos que utilizará el juego, de las habilidades cognitivas que el equipo de expertos en educación planifique que los alumnos pongan de manifiesto al recibirlo y dependerá además, de las características tecnológicas del dispositivo móvil a utilizar durante la experiencia del juego: el tamaño de la pantalla, su resolución, la posibilidad de uso de audio, etc. En [29] se presenta un trabajo que aborda la importancia en la manera en la que se presenta el contenido al alumno en relación a la apropiación del conocimiento como resultado del proceso de aprendizaje. Mientras que en [30] se aborda la problemática de la usabilidad en la presentación de contenido en aplicaciones móviles educativas.
- Diversos estudios [31], [32] y [33] han abordado la problemática de cómo presentar al espacio y de cómo brindar información para acceder a un lugar en el mismo. De acuerdo a [34] los mapas son una buena manera de representar

conocimiento espacial. Sin embargo, cuando se trabaja con dispositivos móviles, se debe tener en cuenta ciertas limitantes, como por ejemplo, el tamaño de la pantalla. Otro aspecto a tener en cuenta, es la edad de las personas que lo usarán, ya que de acuerdo a lo presentado en [35] la manera en la que los niños y los adultos perciben el espacio que los rodea es diferente, debido entre otras cosas, a la diferencia en el conocimiento previo entre ellos. El trabajo presentado por [36] indica que además, la información a representar en los mapas, tienen que ver con la relevancia del propósito con el cual luego serán usados. Si bien estos autores, hablan del espacio en general, estos conceptos pueden ser aplicados en particular al espacio de juego.

Lo antes expuesto, nos permite inferir que un mismo contenido podría ser presentado de diferentes maneras, por ejemplo acorde al perfil de los alumnos. Y acorde a lo presentado en la Sección 3.1, se desprende que es posible reusar la presentación del contenido en diferentes JEMBP.

Por otra parte, en relación a la presentación del espacio de juego, lo más recomendado es el uso de mapas, sin embargo, otro tipo de presentación, podría ser usado, en el caso de ser necesario, por ejemplo un texto descriptivo de cómo llegar a un lugar destacado. Por ejemplo, se puede pensar que a partir de un mapa generado para un JEMBP, el mismo puede ser reutilizado en otro, siempre y cuando, se atiendan a las recomendaciones previamente mencionados en [36]. Con el fin de que la presentación del espacio de juego se pueda reusar, la misma no debe incluir contenido propio de un juego.

Para lograr la generación de un JEMBP, se deben combinar al menos, una presentación de contenido (por cada uno de los contenidos que conforman alguna de las estructuras presentadas en el Sección 3.1) y una presentación del espacio de juego.

En la Figura 3, se muestra la incorporación del aspecto de presentación para uno de los ejemplos presentados en la Figura 2. En particular, para cada contenido, se contempló una sola forma de presentación. Lo mismo, para el espacio de juego.

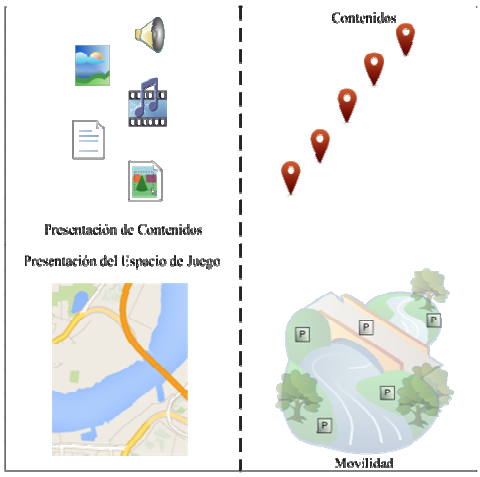


Fig. 3. Aspecto de Presentación

De esta manera, quedan presentados los tres aspectos abordados en este trabajo.

4 Trabajos relacionados

En [37] se menciona el reuso de contenido para aprendizaje móvil, pero no se aborda desde la perspectiva basada en posicionamiento ni contextualizado para espacios de uso replicables, tratando al contenido como objetos de aprendizaje. Por otro lado, el contenido y su presentación están acoplados, es decir, un contenido se reusa con la presentación originalmente definida, sin posibilidad de cambiarla.

En [38] se presenta un framework de aprendizaje móvil, para la adaptación de contenido acorde a las características del dispositivo desde el que se accede, lo que permitiría inferir que ante un mismo contenido, se puede brindar diferentes presentaciones, sin embargo no se mencionan aspectos de reuso de las mismas. Tampoco se menciona la presentación del espacio en el que se usa, pues no aborda el posicionamiento del contenido.

Por otro lado, en [39] se presenta un generador de Juegos Educativos Móviles que permite configurarlos dinámicamente con preguntas de opción múltiple desde un archivo externo. Esto limita el tipo de contenido que se puede brindar al alumno (por ejemplo, no es posible realizar preguntas con respuesta abierta). Para la visualización del espacio de juego, usa la API de Google Maps, lo que imposibilita el reuso con otras presentaciones, por ejemplo, el uso de un mapa esquemático acorde a la edad de los alumnos.

5 Conclusiones y Trabajos Futuro

En este trabajo se presentaron los aspectos de contenido, movilidad y presentación de los JEMBP, de manera desacoplada entre sí, para poder combinarlos, al momento de la creación de los mismos, fomentando de este modo, el reuso de cada aspecto. Este tipo de combinación de aspectos, permite la personalización de JEMBP por parte del equipo multidisciplinario, acordes al objetivo pedagógico del juego, la edad promedio de los alumnos destinatarios, el resultado educativo esperado y a las preferencias de los alumnos.

De acuerdo a [40] se estima que el tiempo para adoptar a los juegos móviles en educación es de dos a tres años. Debido a esto, es la relevancia de trabajar en la temática para lograr buenas prácticas que faciliten la creación y reuso de este tipo de juegos y de acompañar al docente, en el proceso de transformación necesario para adoptar este tipo de desafíos educativos, desde la postura de facilitadores de los alumnos al momento de su uso.

Se trabajará en la especificación de una taxonomía para facilitar la combinación de los aspectos de contenido, movilidad y presentación en la creación de los JEMBP.

Actualmente se está trabajando en la creación de un prototipo de JEMBP basado en un modelo que contempla los aspectos de contenido, movilidad y presentación de

manera desacoplada. Dicho prototipo, se creará empleado la teoría de la actividad [41] para incluir actividades de resolución de problemas educativos. Con el objetivo de realizar una experiencia de uso que permita la evaluación con alumnos, que cursen entre el tercer y cuarto grado del nivel primario.

Referencias

1. Resnick, Lauren B. The 1987 presidential address: Learning in school and out. Educational researcher, vol. 16, no 9, p. 13-54. (1987)
2. Perkins, David N.; Bloberson, Tamar; Salomon, Gavriel. Coparticipando en el conocimiento: la ampliación de la inteligencia humana con las tecnologías inteligentes. CL & E: Comunicación, lenguaje y educación, no 13, p. 6-22. (1992)
3. Salomon, Gavriel. Las diversas influencias de la tecnología en desarrollo de la mente. Infancia y Aprendizaje: Journal for the Study of Education and Development, no 58, p. 143-159. (1992)
4. Quinn, Clark N. Mobile Learning: The Time Is Now. The eLearning Guild's Research. (2012)
5. Laurillard, D. Pedagogical forms of mobile learning: framing research questions. (2007)
6. Johnson, L., Smith, R., Willis, H., Levine, A., Haywood, K.: The 2011 Horizon Report. The New Media Consortium. (2011)
7. Sharples, M., Taylor, J., Vavoula, G.: A theory of learning for the mobile age. In: Andrews, R., Haythornthwaite, C. (eds.). The Sage handbook of learning research, pp. 221-247 (2007)
8. Traxler, J. Learning in a mobile age. *International Journal of Mobile and Blended Learning (IJMBL)*, *1*(1), 1-12. (2009)
9. Brown, S.: Play as an organizing principle: clinical evidence and personal observations. In: Bekoff, M., Byers, J.A. (eds.), Animal play: Evolutionary, comparative, and ecological perspectives, pp. 170-210, Cambridge University Press (1998)
10. Prensky, M.: Digital Game-based Learning. McGraw-Hill, New York (2001)
11. Malone, T.W., Lepper, M.R.: Making learning fun: A taxonomy of intrinsic motivations for learning. In: Snow, R.E., Farr, M.J. (eds.). Aptitude, learning and instruction III: Cognitive and effective process analyses, vol. 3, pp. 223-253 (1987)
12. Susi, T., Johannesson, M., Backlund, P.: Serious games – An overview. School of Humanities and Informatics. University of Skövde, Sweden (2007)
13. George, S., & Serna, A. Introducing mobility in serious games: Enhancing situated and collaborative learning. In *Human-Computer Interaction. Users and Applications*. Pp. 12-20. Springer Berlin Heidelberg. (2011)
14. Schwabe, G., & Göth, C. Mobile learning with a mobile game: design and motivational effects. Journal of computer assisted learning, 21(3), 204-216. (2005)
15. Facer, K., Joiner, R., Stanton, D., Reid, J., Hull, R., & Kirk, D. Savannah: mobile gaming and learning?. Journal of Computer Assisted Learning, 20(6), 399-409.(2004)
16. Raessens, J.F.F.: Playing history. Reflections on mobile and location-based learning. In T. Hug (Ed.), Didactics of microlearning. Concepts, discourses, and examples, pp. 200-217. (2007)

17. Wijers, M., Jonker, V. and Drijvers, P.:MobileMath: exploring mathematics outside the classroom. In ZDM Mathematics Education, Vol. 42, Nº 7, pp. 789-799. (2010)
18. Keskin, N.O. & Metcalf, D. (2011). The current perspectives, theories and practices of mobile learning. The Turkish Online Journal of Educational Technology, 10(2), 202-208. (2011)
19. Traxler, J. Defining mobile learning. In *Proceedings, IADIS international conference on mobile learning, Malta*.(2005)
20. Kadirire, J. Mobile Learning DeMystified . In R. Guy (Ed) *The Evolution of Mobile Teaching and Learning*. California, USA: Informing Science Press. (2009)
21. O'Malley, C., Vavoula, G., Glew, J., Taylor, J., Sharples, M. & Lefrere, P. Guidelines for learning/teaching/tutoring/in a mobile environment. Mobilearn project deliverable. (2003)
22. Ferreira, J. B., Klein, A. Z., Freitas, A., & Schlemmer, E. Mobile Learning: Definition, Uses and Challenges. *Cutting-edge Technologies in Higher Education*, *6*, 47-82. (2013)
23. Traxler, J. Mobile Learning: Shaping the Frontiers of Learning Technologies in Global Context. In *Reshaping Learning* (pp. 237-251). Springer Berlin Heidelberg. (2013)
24. Davidsson, Ola; Peitz, Johan; Björk, Staffan. Game design patterns for mobile games. Project report to Nokia Research Center, Finland. (2004)
25. Schmitz, Birgit; Klemke, Roland; Specht, Marcus. Mobile gaming patterns and their impact on learning outcomes: A literature review. En 21st Century Learning for 21st Century Skills. Springer Berlin Heidelberg. p. 419-424. (2012)
26. Ardito, C., Sintoris, Ch., Raptis, D., Yiannoutsou, N., Avouris, N., Costabile, M.F.: Design Guidelines for Location-based Mobile Games for Learning. In: International Conference on Social Applications for Lifelong Learning, pp. 96-100, Patras, Greece (2010)
27. Lliteras, A.B., Challiol, C. y Gordillo, S.E. Juegos Educativos Móviles Basados en Posicionamiento: Una Guía para su Conceptualización . 41 JAIIO. Agosto de 2012. Facultad de Informática, UNLP. Con referato. In Proceedings of ASSE 2012 Argentine Symposium on Software Engineering. ISSN: 1850-2792, pp. 164-175. (2012)
28. Huang, R., Zhang, H., Li, Y and Yang, J. A Framework of Designing Learning Activities for Mobile Learning. Hybrid Learning Lecture Notes in Computer Science, 2012, Volume 7411/2012. (2012)
29. Yen, J. C., Lee, C. Y., & Chen, I. The effects of image-based concept mapping on the learning outcomes and cognitive processes of mobile learners. British Journal of Educational Technology, 43(2), 307-320. (2012)
30. Ali, A., Ouda, A., & Capretz, L. F. A Conceptual Framework for Measuring the Quality Aspects of Mobile Learning. *Bulletin of the IEEE Technical Committee on Learning Technology*, *14*(4), 31. (2012)
31. Werner, S., Krieg-Brückner, B., Mallot, H. A., Schweizer, K., & Freksa, C. Spatial Cognition: The Role of Landmark, Route, and Survey Knowledge in Human and Robot Navigation1. In Informatik'97 Informatik als Innovationsmotor, pp. 41-50. Springer Berlin Heidelberg. (1997)
32. Klippel, A., Dewey, C., Knauff, M., Richter, K. F., Montello, D. R., Freksa, C., & Loeliger, E. A. Direction concepts in wayfinding assistance systems. In *Workshop on Artificial Intelligence in Mobile Systems*, pp. 1-8. (2004)

33. Tomko, M. *Destination descriptions in urban environments* (Doctoral dissertation, The University of Melbourne). (2007)
34. Zipf, A., & Richter, K. F. Using focus maps to ease map reading. *Künstliche Intelligenz*, *4*(02), 35-37. (2002)
35. Siegel, A.W. & White, S.H. The development of spatial representati-ons of large-scale environments. In H.W. Reese (ed.) Advances in Child Development and Behavior , pp. 9–55, New York: Academic Press. (1975)
36. Tenbrink, T. Relevance in spatial navigation and communication. In Spatial Cognition VIII , pp. 358-377. Springer Berlin Heidelberg. (2012)
37. Tabuenca, B., Drachsler, H., Ternier, S., & Specht, M. OER in the Mobile Era: Content Repositories' Features for Mobile Devices and Future Trends. (2012)
38. Madjarov, I., & Boucelma, O. XESOP: a content-adaptive m-learning environment. In *21st Century Learning for 21st Century Skills*, pp. 531-536. Springer Berlin Heidelberg. (2012)
39. Schmitz, B., Klemke, R., Specht, M., Hoffmann, M., & Klamma, R. Developing a Mobile Game Environment to Support Disadvantaged Learners. In *Advanced Learning Technologies (ICALT), 2012 IEEE 12th International Conference on* , pp. 223-227. IEEE. (2012)
40. Johnson, L., Adams, S., & Cummins, M. NMC horizon report: 2012 higher education edition. (2012)
41. Engeström, Y. Expansive learning. *Contemporary Theories of Learning*, pp.53-73. (1987)

Characterization of University Drop-Out at UNRN Using Data Mining. A Study Case

Sonia Formia [1], Laura Lanzarini [2] and Waldo Hasperué [2,3]
[1]Applied Computer Science Laboratory – LIA Bachelor's Degree in Systems, UNRN - Atlantic Coast Delegation
[2] Institute of Research in Computer Science – LIDI School of Computer Science. UNLP
[3] CONICET scholarship
sformia@unrn.edu.ar, {laural, whasperue}@lidi.info.unlp.edu.ar

Abstract. At the National University of Río Negro (UNRN), and its Atlantic Coast Delegation in particular, it is an increasing concern for the courses corresponding to the Bachelor's Degree in Systems, the drop-out and crumbling rates observed in the first four years of the Institution. This paper describes the process of identifying the most relevant features of the problem through which, using Data Mining (DM) techniques, a college drop-out model can be obtained for the academic unit mentioned above. In order to identify the most relevant features, after processing the data we will analyze attribute projections for the expected classes or responses. The results of its application to the student data from the courses of the UNRN have been satisfactory, which allows making some recommendations aimed at reducing the percentage of students who drop put from their courses.

Keywords: Attribute Selection. Attribute Projection. Data Mining. University Drop-Out.

1 Introduction

The organization being studied is the National University of Río Negro, which was founded in 2008 and started its graduate programs in 2009. There are currently a total of 60 graduate programs. From the very beginning, both the authorities and the educators working in the various programs have been concerned about the high drop-out and crumbling rate observed, despite the short life of the Institution. The main purpose is being able to determine potential academic failure situations in advance so as to apply measures aimed at minimizing the problem.

On the road to achieving the ultimate goal, predicting drop-out, other goals may be found that contribute non-trivial, useful information for the decision-making process, such as describing or characterizing UNRN students by means of profiles that help guide the implementation of measures at those levels where they can have the greatest positive effect.

Several authors have proposed solving this problem through various approaches, both involving student recruitment and drop-out detection and analysis, as well as for assessing the duration of the program [1] [2] [3] [4] [5] [6] [7]. Recently,

environments aimed at facilitating the application of DM techniques in educational contexts have been developed [8].

This paper is part of what is known as the Knowledge Discovery in Databases (KDD) process, whose purpose is the automatic discovery of patterns present in available information without specifying any hypotheses beforehand. Its application requires the identification, based on the problem to be solved, of the information on which work is going to be done, as well as the desired type of model to be obtained. The latter strongly affects the technique to be used.

The available information is obtained from the SIU-Guarani system of the UNRN. That is, for each student there is a significant number of features pertaining to their personal, academic and work situations. Selecting the appropriate features from this data set to build a model can be a challenge. There is a proportional relation between the number of attributes to be used and the complexity of the model to be obtained. For these reasons, the early discovery of the most relevant attributes is desirable in order to simplify the model and reduce the time required to obtain it.

This paper is organized as follows: Section 2 describes the pre-processing stage carried out on the original data, Section 3 details how these were considered, Section 4 details other strategies previously used to solve this problem, Section 5 describes the attribute selection process based on projections, Section 6 presents a model based on the selected attributes, and Section 7 presents the conclusions drawn.

2 Preparing UNRN Data

Before applying a specific DM technique, the data had to be verified to avoid inconsistencies. This stage was guided by the data preparation methodologies surveyed in the literature and previous experience in the field.

Those attributes with an excessive number of missing data were deleted, any outlying values were cleaned, constant and redundant attributes were removed, and generalization was used to transform high-cardinality attributes. Non-generalizable attributes were deleted, the cardinality of some attributes was reduced by using more general categories, new attributes were built by means of summarization operations, attributes were discretized based on algorithm requirements, and range normalizations were performed. Finally, a state attribute was defined that differentiates between students who have dropped out (after one year with no academic activity) and those who are regular students.

One of the main problems in DM is identifying a representative set of appropriate features to build a model for the task at hand. Problems with a large number of dimensions, limited numbers of available examples, and a lot of redundant or irrelevant information are hard to handle [9]. This study case clearly has these features: the SIU-Guarani provides a significant number of attributes for the students, and the number of examples is limited due to the short life of the UNRN – only information from the last 4 academic years is available. The original (or initial) database view used in this paper is formed by 11,102 students, each with 110 surveyed attributes.

3 Approach

As a first measure, to understand the data available and describe the domain of the problem, it was decided that the target set of students being studied, i.e., the records corresponding to those students who dropped out, would be characterized. This would allow selecting the most relevant features for drop-outs. Among the descriptive tasks provided by data mining, clustering is one of the most frequently used; its purpose is obtaining groups or sets within the examples, so that the elements assigned to the same group are similar [10].

In our study case, the information available includes demographic, economic, social, family, and academic data of the students. By means of the k-means clustering algorithm, drop-outs were segmented into groups. The tests carried out showed the true dimension of the problem. The large number of attributes involved did not allow finding a set of descriptive clusters for the input data.

At this point, DM tools had to be used to guide the selection of a subset of features (attributes) that are relevant for the problem.

4 Previous Work

The authors in [19] used the input data set formed by drop-outs and transformed the feature space by means of two completely different processes: one of the *wrapper* type and another one of the filter type.

Wrapper processes classify the attributes selected based on the performance of the model that can be built from them [11]. There are various ways to do this. The authors in [19] used a selection process of the *Selection Forward* type. This technique starts the feature search procedure by assessing all subsets of attributes formed by a single attribute, then finds the best subset of two attributes, then a subset of three attributes, and so forth until the best subset of features is found. To validate the sets of features, the performance of a given learning model is considered. In this case, the k-means method was used to group the available information. The use of an inductive algorithm is what makes this method a *wrapper*-type process.

The second selection method that was implemented is focused in the genetic selection of the features [12]. The genetic algorithm carries out a heuristic search that minimizes the natural evolution process. For the assessment, it uses the CFS (*Correlation-based* Features *Selection*) method, which creates a filter based on the performance measured for the set of features. It assessed the value of a subset of attributes considering the predictive ability of each feature together with the redundancy degree among them, and giving preference to subsets of attributes that are highly correlated with the class but whose inter-correlation is low [13].

Once both methods were implemented on the input data, it could be seen that they both yielded similar subsets of attributes (see Table 1).

Table 1. List of attributes selected by the wrapper and genetic methods

Description		Attribute Name	Selected By	
			Wrapper	Genetic
The student is single		estado_civil = single	YES	YES
The father of the student is alive		padre_vive = YES	YES	YES
		situacion_laboral_padre	YES	YES
Internet access at home		alu_tec_int	YES	NO
The student acknowledges that he/she needs a scholarship		alu_beca = needs scholarship	NO	YES
Student's current employment	Employment situation	alu_trab_sitimp	YES	YES
	Related to course being taken	rel_trab_carrera	YES	YES
	Monthly income	alu_trab_remmon	YES	YES
Delegation where the student takes classes		Delegation	YES	YES
Place of birth		lugar_nacimiento	YES	YES
High School Graduation Year		anio_egreso_sec	NO	YES
The student is planning to work in the future	Type of work	alu_trab_futtip	YES	YES
	Time	alu_trab_futhor	YES	NO
Year of Birth		anio_nacim	YES	NO
Number of dependent family members		cant_fami_cargo	YES	YES
Student's Number of Children		cant_hijos_alum	YES	YES

In order to check the validity of the attributes selected by the methods described, all drop-out student records were re-grouped based on the attributes selected by the wrapper method and supported by the other algorithms presented. The groups resulting after this run are compared with those obtained with the previous one, and it was observed that less than 10% of the students had been assigned to a different group, indicating that the clustering criterion remains the same despite the use of a smaller set of features.

Then, the same clustering algorithm was applied to non drop-outs and a similar segmentation was obtained for the main attributes to that of drop-outs, which supports the use of the selected attributes in predictive drop-out algorithms.

5 SOAP: Selection of Attributes by Projection

This section describes how to select the most representative attributes using the SOAP (*Selection of Attributes by Projection*) method [15]. Unlike those used in the previous section, this is a filter that measures attributes and establishes a deterministic ranking to reduce computation time.

This method adds a new criterion to measure the significance of an attribute within a supervised learning context: it uses the number of label changes. This value is calculated by analyzing dataset element projections on each attribute. Thus, attributes can be sorted by significance when establishing their class (in this case, whether the student drops out from the program or not).

For attribute selection, a measure is used that is based on a single value: NLC (*Number of Label Changes*), relating each attribute with the label used for classification. The value of NLC is calculated by projecting the examples on the axis that corresponds to this attribute (i.e., by sorting all examples by the attribute in question), and then the axis is run from its origin up to the highest attribute value and the number of label changes that occur is counted:

SOAP Algorithm.
Input: E–data set (m examples, n attributes)
Output: R–attribute ranking

```
R ← {}
for i = 1 to n do
  Sort E by attribute Xi
  NLCi ← CountChanges(E; Xi)
end for
R (attribute ranking based on NLC)
```

Function `CountChanges(E; Xi)`
Input: *E–data set (m examples, n attributes), X_i–attribute to process*
Output: *Changes–Number of label changes*

```
R ← {}      Changes ← 0
for j = 1 to m do
  if xj;i ε Multiple_Sorted_Sequence
       Changes← Changes + SeeChangesForSameValue()
  else
       if lab(ej) <> lab(ej+1) then
            Changes ← Changes + 1
       end if
  end if
end for
return(Changes)
```

In `CountChanges`, it should be noted that, when sorting attribute values, there might be repeating values. This will lead to a `Multiple_Sorted_Sequence` (MSS), where the value of the attribute is the same for all examples but the labels assigned to each of them will not necessarily be the same. In this case, the function `SeeChangesForSameValue()` is applied, which calculates the number of corresponding changes as follows: If all examples have the same label, then the returned number of changes is zero; otherwise, first we need to know if there is a majority class within the group of examples that share the same attribute value. If there is no majority class, the number of changes is the length of the MSS minus 1. If there is a majority class, the number of changes is the number of elements in the MSS minus the number of elements of the majority class.

Table 2. First attributes in the ranking as per SOAP

NLC	Attribute
2186	alu_trab_remmon
2186	Alu_trab_sitimp
2252	rel_trab_carrera
3044	alu_trab_futtip
3571	alu_trab_futhor
3987	anio_egreso_sec
4004	hora_sem_trab_alum
4063	anio_nacim
5070	alu_otestsup_uni
5394	Sit_laboral_madre
5734	cant_hijos_alum
5747	Sit_laboral_padre

5.1 Application of the SOAP Algorithm to the Study Case

The algorithm described above was implemented and applied to the study case. Only those records corresponding to students who dropped out from the program were considered, and were previously grouped by k-means to assign different labels to them. As in the processes described in the previous section, 5 groups were used (k=5).

The results of applying the SOAP algorithm to drop-out student data, segmented into five groups, yield a full attribute ranking, where each attribute receives an NLC value. Top ranking values (lower NLC values, i.e., attributes that project a lower number of label changes) are shown in Table 2, sorted by NLC value.

If the attributes appearing at the top of the ranking are observed, it can be seen that those attributes that are related to student employment are predominant, together with those that describe student age and family responsibilities. These attributes were in general present in all attribute lists obtained with the feature selection algorithms that were implemented first (Table 1).

It can then be concluded that the most relevant features mainly include attributes pertaining to student age, work load, and family responsibilities.

Given the scarcity of examples available for research (due to the short life of the UNRN) and the large number of attributes at the beginning of the tasks, it can be inferred that, for the time being and until there are more examples to feed other algorithms, this group of personal and employment attributes must be accepted as those describing drop-out and non drop-out students. With this idea, a predictive algorithm, in this case a decision tree, can be applied to all of the examples available using a group of attributes that are high in the SOAP ranking to classify students as drop-outs or non drop-outs. Thus, the tree obtained is much simpler than the one that can be achieved with the original set of attributes.

Table 3: Performance of the tree built with the selected attributes (Table 2) using the C4.5 method and considering a confidence threshold of 0.25 and a minimum number of 2 elements per leaf.

Accuracy:69.84%	**True Drop-out**	**True Taking courses**	**Class precision**
Pred Drop-out	5174	2047	**71.65%**
Pred Taking courses	1388	2414	**63.49%**
Class Recall	**78.85%**	**54.11%**	

6 Predictive Drop-Out Model

Tests were carried out with algorithm C4.5 [14] using the top nine attributes in the SOAP ranking (Table 2), except for "anio_egreso_sec," which corresponds to the year the student graduated from high school. This is because the attribute "anio_nacim," representing the year the student was born, is already included, and both attributes are closely related to each other. Table 3 shows the performance of the model obtained.

It can be seen that the length of the list of attributes used is 56.25% of the list of attributes in Table 1 (9/16 = 0.5625). The attributes that were selected allow building a model that can successfully predict 71.65% of drop-out cases. The success rate is lower when predicting if the student is still taking courses.

Figure 2 shows a pruned version of the resulting tree considering a minimum number of 10 elements per leaf.

7 Conclusions

This paper presents the application of a feature selection method based on projections that can operate on nominal and numeric attributes in a supervised manner. From the ranking it establishes for attributes, a cut-off point can be determined to identify those that are the most representative. In this case, its application allowed reducing the original list (Table 1) by more than 40% (the first 9 attributes in Table 2).

Table 2 shows that the most relevant attributes are those pertaining to student employment status, both regarding current employment and future employment plans.

As preliminary product, clear guidelines can be obtained to guide the measures to be implemented to reduce student drop out rates at the UNRN: it is clear that student employment variables have a significant effect on their likelihood to continue with their studies, so any actions that specifically target this issue, such as a larger number of scholarships granted, could be a road to follow.

Beyond these conclusions, a predictive model was proposed that can be improved in time as more examples are added to the data set.

```
anio_nacim <= 1989: Drop-out (7614/2538)
anio_nacim > 1989
|  anio_nacim <= 1992
|  |  alu_otestsup_uni = "Y"
|  |  |  Alu_trab_sitimp = "Employed": Taking courses (65/21)
|  |  |  Alu_trab_sitimp = "does not work": Taking courses (257/90)
|  |  |  Alu_trab_sitimp = "Self-employed": Drop-out (11/4)
|  |  alu_otestsup_uni = "N"
|  |  |  Alu_trab_futtip = "Worker or employee (salary)"
|  |  |  |  Alu_trab_futhor = "10 to 20 hs": Taking courses (93/42)
|  |  |  |  Alu_trab_futhor = "Will not work": Taking courses (7/2)
|  |  |  |  Alu_trab_futhor = "> 35 hs"
|  |  |  |  |  Hora_sem_trab_alum = "Not reported": Drop-out (33/13)
|  |  |  |  |  Hora_sem_trab_alum = "21 to 35 hs": Drop-out (0)
|  |  |  |  |  Hora_sem_trab_alum = "> 20 hs": Taking courses (4/1)
|  |  |  |  |  Hora_sem_trab_alum = "> 36 hs": Taking courses (1)
|  |  |  |  |  Hora_sem_trab_alum = "does not work": Drop-out (48/17)
|  |  |  |  Alu_trab_futhor = "< 10 hours": Drop-out (128/51)
|  |  |  |  Alu_trab_futhor = "20 to 35 hs"
|  |  |  |  |  Hora_sem_trab_alum = "Not reported": Drop-out (25/10)
|  |  |  |  |  Hora_sem_trab_alum = "21 to 35 hs": Taking courses (3/1)
|  |  |  |  |  Hora_sem_trab_alum = "< 20 hs": Drop-out (2/1)
|  |  |  |  |  Hora_sem_trab_alum = "> 36 hs": Drop-out (36/15)
|  |  |  |  |  Hora_sem_trab_alum = does not work: Taking courses (11/3)
|  |  |  Alu_trab_futtip = "Will not work": Taking courses (1358/613)
|  |  |  Alu_trab_futtip = "Self-employed"
|  |  |  |  Alu_trab_remmon = "1200 to 2000$": Drop-out (9/4)
|  |  |  |  Alu_trab_remmon = "> 3000$": Taking courses (2)
|  |  |  |  Alu_trab_remmon = "does not work": Drop-out (35/16)
|  |  |  |  Alu_trab_remmon = "< 1200$": Taking courses (33/14)
|  |  |  |  Alu_trab_remmon = "2000 to 3000$": Drop-out (2/1)
|  |  |  Alu_trab_futtip = "Does not know"
|  |  |  |  Alu_trab_futhor = "10 to 20 hs": Drop-out (81/28)
|  |  |  |  Alu_trab_futhor = "Will not work": Drop-out (107/52)
|  |  |  |  Alu_trab_futhor = "> 35": Drop-out (6/3)
|  |  |  |  Alu_trab_futhor = "< 10 hs": Taking courses (138/66)
|  |  |  |  Alu_trab_futhor= "20 to 35 hs": Drop-out (29/9)
|  anio_nacim > "1992"
|  |  Hora_sem_trab_alum = "Not reported": Taking courses (827/278)
|  |  Hora_sem_trab_alum = "21 to 35 hs": Taking courses (13/5)
|  |  Hora_sem_trab_alum = "< 20 hs": Taking courses (26/9)
|  |  Hora_sem_trab_alum = "> 36 hs": Drop-out (10/4)
|  |  Hora_sem_trab_alum = "does not work": Drop-out (9/2)
```

Fig. 2. Decision tree to determine drop-outs

References

1. La Red Martínez, D. L., Acosta, J. C., Cutro, L. A., Uribe, V. E., and Rambo, A. R. (2009). Data warehouse y data mining aplicados al estudio del rendimiento académico y de perfiles de alumnos. In XII Workshop de Investigadores en Ciencias de la Computación – CACIC 2010, pages 162–166.
2. Luo, Q. (2008). Advancing knowledge discovery and data mining. In Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on.
3. Alcover, R., Benlloch, J., Blesa, P., Calduch, M. A., Celma, M., Ferri, C., Hernández Orallo, J., Iniesta, L.,Más, J., Ramírez Quintana, M. J., Robles, A., Valiente, J. M., Vicent, M. J., and Zúnica, L. R. (2007). Análisis del rendimiento académico en los estudios de informática de la universidad politécnica de valencia aplicando técnicas de minería de datos. Technical report, Universidad Politécnica de Valencia.
4. La Red Martínez, D. L., Acosta, J. C., Cutro, L. A., Uribe, V. E., and Rambo, A. R. (2009). Data warehouse y data mining aplicados al estudio del rendimiento académico y de perfiles de alumnos. In XII Workshop de Investigadores en Ciencias de la Computación – CACIC 2010, pages 162–166.
5. Luo, Q. (2008). Advancing knowledge discovery and data mining. In Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on.
6. Alcover, R., Benlloch, J., Blesa, P., Calduch, M. A., Celma, M., Ferri, C., Hernández Orallo, J., Iniesta, L.,Más, J., Ramírez Quintana, M. J., Robles, A., Valiente, J. M., Vicent, M. J., and Zúnica, L. R. (2007). Análisis del rendimiento académico en los estudios de informática de la universidad politécnica de valencia aplicando técnicas de minería de datos. Technical report, Universidad Politécnica de Valencia.
7. Valero, S. and Salvador, A. (2009). Predicción de la deserción escolar usando técnicas de minería de datos. In Simposio Internacional en Sistemas Telemáticos y Organizaciones Inteligentes SITOI 2009, pages 332–340.
8. Rodallegas, E., Torres, A., Gaona, B., Gastelloú, E., Lezama, R., and Valero, S. (2010). Modelo predictivo para la determinación de causas de reprobación mediante minería de datos. In II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje – CcITA 2010, pages 48–55.
9. Valero, S., Salvador, A., and García, M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. In II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje – CcITA 2010, pages 33–39.
10. Wang, J., Lu, Z., Wu, W., and Li, Y. (2012). The application of data mining technology based on teaching information. In Computer Science Education (ICCSE), 2012 7th International Conference on, pages 652 –657.
11. Ngo, L., Dantuluri, V., Stealey, M., Ahalt, S., and Apon, A. (2012). An architecture for mining and visualization of u.s. higher educational data. In Proceedings of the 2012 Ninth International Conference on Information

Technology - New Generations, ITNG ’12, pages 783–789, Washington, DC, USA. IEEE Computer Society.

12. Hernández Orallo, J., Ramírez Quintana, M., and Ferri Ramírez, C. (2004). Introducción a la Minería de Datos. Ed. Pearson.
13. Witten, I. H. and Frank, E. (2011). Data Mining: Practical Machine Learning.Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco, CA, 3th edition.
14. Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artif. Intell., 97(1-2):273–324.
15. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st ed.
16. Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand.
17. Quinlan, R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
18. Sanchez, Roberto Ruiz. Heurísticas de selección de atributos para datos de gran dimensionalidad. Doctoral Thesis, University of Seville.
19. Formia S, Lanzarini L. Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios. VIII Congreso de Tecnología en Educación y Educación en Tecnología 2013. Sgo.del Estero. June 2013.

Arte y TIC: Experiencias iniciales con herramientas de software en la formación de Licenciados en Artes Combinadas

Fernández M.[1,2], Barrios W.[2] Godoy M.V.[2] y Gendin G.[1]
[1]Facultad de Artes, Diseño y Ciencias de la Cultura
[2]Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura-
Universidad Nacional del Nordeste, Corrientes, Argentina.

mirtagf@hotmail.com, mvgg2001@yahoo.com, ggendin@yahoo.com, waltergbarrios@yahoo.com.ar

Abstract: Se presenta una experiencia educativa en un grupo de estudiantes universitarios que participaron de una propuesta pedagógica innovadora, donde el objetivo principal fue crear un ámbito de reflexión, experimentación, integración y vínculos de la ciencia, el arte y la tecnología. Integrada por tres ejes de la currícula: en el primero se abordó la evolución histórica y la relación entre las tres disciplinas, antes mencionadas; en el segundo, se trabajó el proceso creativo, a partir de la construcción de un artefacto artístico en relación a los medios analógicos; en la tercera se introdujo conceptos de medios digitales. Acompañados de la bibliografía propuesta, las actividades se sustentaron en la simulación, a fin de comprender los fenómenos y las leyes físicas que presenta la ciencia, su implicancia en el arte y la tecnología. Al finalizar, se extraen conclusiones al respecto.

Palabras clave: Tecnologías de la Información y la Comunicación, Enseñanza en Artes, Configuraciones Multidisciplinares, Imagen-Sonido-Espacio-Tiempo.

1 Introducción

Las Artes Combinadas surgen como producto de la expansión de los campos y fronteras de las expresiones artísticas a lo largo del siglo XX [1].
En estos procesos se han establecido diálogos entre diferentes disciplinas tradicionales como la literatura, la pintura, el dibujo, la escultura, la música, el teatro, la danza, la fotografía y el cine. De esos entrecruzamientos de lenguajes específicos se han constituido categorías y prácticas transversales a los mismos, estableciéndose múltiples modalidades de intercambio entre las mismas, que articulan una fuerte fundamentación teórica y una práctica intensiva.

Es a partir de ello, y respondiendo a áreas de vacancia en la formación y ejercicio profesional de la región que se inició en 2012, a través de la nueva carrera de grado de Licenciatura en Artes Combinadas perteneciente a la FADYCC[1] de la UNNE[2] (Chaco) y en particular, en este trabajo se hará referencia a la propuesta educativa de la asignatura "Introducción a la Tecnología Aplicada al Arte".

1.1 TIC y "nuevos medios"

Como producto de la investigación y evolución de las Tecnologías de la Información y Comunicación (TIC), hoy es posible observar el cruce necesario de estas, con las disciplinas consideradas específicas, así como por las diversas expresiones del arte [2]. De ello, surgen números abordajes y posturas [3], [4], [5], [6] y [7].

Estas manifestaciones tienen su génesis de producción de arte por medio de un ordenador y más extensivo aún, conectados en la red; experiencias que en su generalidad han sido denominadas como "nuevos medios", un término cuestionado, debido a lo amplio e indeterminado que resulta el concepto de "nuevo".

Según Manovich [5], la comprensión popular de *nuevos medios* los identifica con el uso del ordenador para la distribución y la exhibición de contenidos, más que con la producción. Un ejemplo de ello son los sitios web y los libros electrónicos, se consideran nuevos medios, en tanto, los que se distribuyen en papel, no [7].

Desde esta perspectiva, el autor llama a la reflexión: ¿Se debe aceptar ésta definición? Al respecto, alude que "no hay motivo para privilegiar el ordenador como aparato de exhibición y distribución por encima de su uso como herramienta de producción o como dispositivo de almacenamiento", y lo justifica retrotrayéndose a dos recorridos históricamente separados, como son las tecnologías informáticas y mediáticas.

1.2 Divergencia de los medios

En 1839, Louis Daguerre presenta la descripción formal de un nuevo proceso de reproducción, llamado *daguerrotipo* [8], por el cual se obtiene una imagen en positivo a partir de una placa de cobre recubierta de yoduro de plata.

En 1833, Charles Babbage construye un aparato que llamó la *máquina analítica*, y que ya contenía la mayoría de las principales características del ordenador digital moderno [9], usaba tarjetas perforadas para su funcionamiento. Esta máquina no prosperó en su creación y de ello se postulan diversas conjeturas [10] y [11].

[1] http://www.artes.unne.edu.ar/Artes-Combinadas.html
[2] http://www.unne.edu.ar/

Coincidentemente, Babbage toma la idea de una máquina programada hacia 1800; un telar, inventado por J. M. Jacquard. Respecto a ello, Ada Augusta, la primera programadora informática indicó: "La máquina analítica teje patrones algebraicos igual que el telar de Jacquard teje flores y hojas" [12].

Mientras, que la invención del daguerrotipo impactó en la sociedad de manera inmediata, el impacto del ordenador no llegaba.

Ambas trayectorias discurrían en paralelo. A lo largo del siglo XIX y a comienzos del XX se desarrollaron numerosos tabuladores y calculadoras mecánicas y eléctricas. Y por su parte, al auge de los medios modernos que permiten guardar imágenes, secuencias de imágenes, sonidos y texto, por medio de diferentes formas materiales: placas fotográficas, películas, discos, etcétera.

En la década de 1890, los medios modernos pusieron las fotografías en movimiento. El primer estudio *cinematográfico*, de Edison, comenzó a realizar cortos de treinta segundos. Más tarde, los hermanos Lumiére mostraron su nuevo híbrido de cámara y proyector cinematógrafo. Esa misma época fue crucial para la informática. Herman Hollerith, consolida las ideas de telar de Jacquard y la máquina analítica de Babbage, patentando una máquina electromecánica de información que utilizaba tarjetas perforadas, para el procesamiento del Censo de Estados Unidos.

1.3 Convergencia de los medios

Diversos autores, concuerdan que la década clave en la historia de los medios y de la informática es 1930 [3] y [5]; en ella se desarrolla un moderno ordenador digital. El matemático británico Alan Turing en su artículo titulado "Sobre los números computables", proporcionaba una descripción teórica de la *máquina universal de Turing*. La misma, funcionaba a base de leer y escribir números en una cinta sin fin que a cada paso avanzaba recuperando la siguiente orden, leyendo los datos o escribiendo el resultado y guardando semejanza con el de un proyector de cine.

Una nueva coincidencia se pone de manifiesto: *cinematógrafo*, significa "movimiento escrito", es decir, la esencia del cine es registrar y guardar datos visibles en una forma material. Una cámara de cine registra unos datos sobre película y el proyector los lee uno por uno. Este aparato se parece al ordenador en un aspecto esencial: el programa y los datos del ordenador también se tienen que guardar en algún soporte.

El desarrollo de un medio de almacenamiento adecuado y de un método para codificar los datos representan partes importables de la prehistoria tanto del cine como del ordenador. El cine utilizó imágenes discretas, que quedaban registradas en una tira de celuloide; el ordenador, adoptó un almacenamiento electrónico sobre código binario.

1.4 El encuentro definitivo de los medios

Las historias de los medios y la de la informática se entrelazaron más aun, cuando el ingeniero alemán Konrad Zuse, construye el primer ordenador digital; empleando cinta perforada para controlar los programas del ordenador. La cinta era, en realidad, trozos de descartes de película cinematográfica [12].

Así el discurso, el sentido y la emoción que contuviera esa secuencia cinematográfica, habían quedado anulados por su nueva función como soporte de datos (materia prima, insumo, etc.). El código icónico del cine queda descartado en favor del binario, más eficiente. El cine se vuelve dependiente del ordenador [5].

Este encuentro cambia la identidad tanto de los medios como del propio ordenador, que deja de ser sólo una calculadora, un mecanismo de control de un dispositivo de comunicaciones, para convertirse en un *procesador de medios*. Todo ello, tiene explicaciones inherentes, en fenómenos y leyes físicas que presenta la ciencia.

1.5 Caracterización del espacio curricular

La Licenciatura en Artes Combinadas brinda un campo de experimentación continuo para la producción e investigación, siendo las TIC un componente que atraviesa las distintas expresiones y perspectivas artísticas. La carrera se inicia en 2011, con un importante número de ingresantes; en la cohorte 2012 se inscribieron 220 estudiantes.

El desafío pedagógico que representa particularmente, la asignatura de primer año "Introducción a la Tecnología Aplicada al Arte", *es crear un ámbito de reflexión, experimentación e integración*; donde sea posible indagar las relaciones, próximas y a la vez problemáticas, entre las ciencias y las artes, mediadas por las nuevas tecnologías.

Para atender con éxito estos procesos de enseñanza-aprendizaje emergentes, se requieren nuevas configuraciones multidisciplinarias [13], para su abordaje. Es así, que una de las estrategias adoptadas fue la integración del cuerpo docente: un Licenciado en Música con formación en Arte multimedial, un Ingeniero Civil, una Arquitecta y una Licenciada en Sistemas de Información, sumando aportes y visiones cada uno desde su disciplina particular, en función de lograr el objetivo central.

La modalidad de dictado es presencial, con una carga horaria total de 96 horas. Se cuenta con un Laboratorio Informático, con capacidad para 40 alumnos, por lo que la totalidad de los estudiantes se dividen en comisiones, para el desarrollo de las clases de informática.

Los objetivos específicos de la asignatura, se exponen a continuación:

- Lograr procesos de enseñanza-aprendizaje integradores en relación con los demás niveles o asignaturas del plan de estudios.
- Estimular en los alumnos, inquietudes para el desarrollo de tareas de investigación.

- Promover la formación de profesionales en condiciones de adaptarse a los constantes cambios del mercado artístico laboral.

3 Marco Metodológico

Se parte del presupuesto: la tecnología no es el arte en sí, sino una herramienta que mutó y acompañó a lo largo del tiempo los múltiples discursos artísticos en las distintas épocas. Llegando a introducir elementos teóricos y funcionales de ciencias tales como la física cuántica, la genética, la biología, las TIC, tal como lo presentan artistas multimediales de referencia, como el Grupo Biopus [14].

Para el desarrollo de las temáticas planteadas:

- Se delimitan tres ejes principales que sustentan los contenidos teóricos del programa analítico; denominando Eje Temático 1(ET1), Eje Temático 2 (ET2) y Eje Temático 3 (ET3).
- Se diseñan las actividades áulicas y extra-áulicas de los ejes.
- Se describen los criterios evaluativos, de las composiciones logradas por los alumnos en dichos ejes y se exponen algunos trabajos a modo de ejemplos.

4 Resultados

Se presentan los ejes, y en este artículo se describen con mayor detalle los **ET2** y **ET3**, a fin de reflejar el uso de material o herramienta (analógico o digital) y la importancia del discurso del relato en el arte, por sobre ellos.

4.1 Eje Temático 1

El **ET1** se centra en la reflexión sobre los nuevos medios tecnológicos, estableciendo el vínculo entre arte, ciencia y tecnología. Siguiendo una cronología que se extiende desde las expresiones artísticas primitivas hasta las contemporáneas, su desarrollo individual y convergencia histórica. Asimismo, se insta a repensar las relaciones o interrelaciones entre arte, medios audiovisuales, medios digitales y TIC, promoviendo la lectura de autores que abordan estas distintas perspectivas [5].

Sustentados en el planteo o marco teórico del espacio curricular, se introducen conceptos, del pasaje de lo analógico a lo digital, es decir de continuo a discreto, elementos fundamentales para comprender el paradigma digital; así como nociones de óptica y mecánica.

4.2 Eje Temático 2

En el ET2 se abordó el proceso creativo, vinculándolo al concepto Imagen- Sonido-Espacio-Tiempo. El nexo se plasmó, en un trabajo práctico con sus respectivas etapas de producción; con la construcción de un dispositivo (objeto o artefacto) realizado con tecnología *analógica* y materiales cotidianos; a partir de una idea generadora y la comprensión de un texto seleccionado, una obra teatral llamada “Complejo de Edipo”.

En relación a conceptos específicos, se abordaron:

- Principios básicos de la dimensión espacio-temporal.
- Principios básicos ópticos
- Principios básicos sonoros

4.2.1 Actividad Práctica ET2

Para la materialización y desarrollo de la unidad, se propuso la construcción de un artefacto que antecedió al cine, que genere imagen y sonido (Ej: linternas mágicas, zootropo, praxinoscopio, kinetoscopio, Arte Cinético y objetos sonoros de la Corriente Futurista Italiana).

Esta experiencia plantea el simulacro de un proceso creativo, al que el alumno se enfrentará en el momento de realizar una obra de arte interdisciplinaria.

Se trata de una máquina del pre-cine extraído del **ET1**, un Zootropo; que busca producir la ilusión de imagen en movimiento, resultado de conceptos ópticos y mecánicos. Un Zootropo, como lo presenta la **Fig. 1**, es un tambor con ranuras por las que se visualizan imágenes secuenciales, impresas sobre una tira de papel colocada en el interior del tambor.

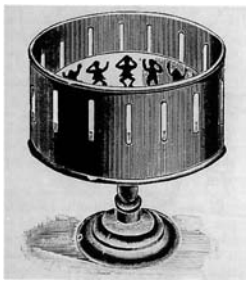


Fig. 1: Un Zootropo es considerado un Objeto o Juguete del pre-cine.

Se construyó en el aula (grupalmente) una tira de cuadros de imágenes (fotogramas) secuenciales, a modo de ejemplo. Como actividad extra-áulica y apelando a la creatividad espontánea grupal, se propuso la intervención del artefacto y la diagramación de las 5 tiras con 10 fotogramas, a partir de la selección de 5 conceptos abstractos, presentes en el texto acordado.

4.2.2 Evaluación de ET2

Se buscó comprender al proceso creativo como un sistema de etapas abierto, que recibe retroalimentación y seguimiento de los docentes (feedback) para la resolución de un "problema". Se consideró importante el espíritu interdisciplinario en el trabajo grupal y el discurso por sobre el material y su funcionamiento. La actividad se materializó desde la idea generadora hasta su ejecución.

4.3 Eje Temático 3

En el **ET3** se abordó mediante clases teóricas y prácticas, la **incidencia del ordenador y el software en las artes** [7]**.** Se desarrolló la función de las herramientas

informática, el código binario, los algoritmos de codificación y tipos de aplicaciones, desde un punto de vista *digital*. En la plataforma informática, y continuando la narrativa del Zootropo, se presentan conceptos de imagen en movimiento, como inicios del cine.

En relación a conceptos específicos, en este eje se desarrollaron:

- Nociones de imagen y sonido digital.
- El Algoritmo, características y funciones.
- Las herramientas del Software y la interacción.
- Línea de Tiempo, Fotogramas, símbolos, capas y acciones.

4.3.1 Actividad Práctica ET3

Para lograr la introducción de conceptos de fotogramas, se visualizó la composición y descomposición de imágenes percibidas de forma continua, en imágenes discretas como se presenta en la **Fig.2.**

Fig. 2. Primeras imágenes del pre-cine.

Se retomó el texto trabajado en la **ET2** y se digitalizaron los fotogramas del Zootropo, en la plataforma.

4.3.2 Evaluación de ET3

En el trabajo práctico final, se incorporaron las imágenes digitalizadas en un collage interactivo. Con ello se reforzó el concepto de interactividad sobre una plataforma informática concreta. Como resultado del **ET3** se obtuvieron trabajos que se muestran en la **Fig. 3, Fig. 4** y **Fig. 5**.

Fig. 3. Un concepto representado en este collage es “La búsqueda de la verdad”

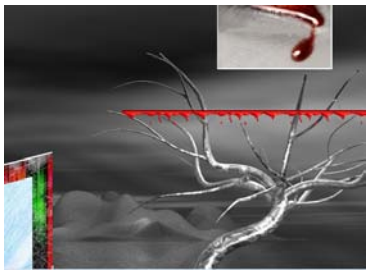


Fig. 4: Un concepto seleccionado para este trabajo es “La Vida”.

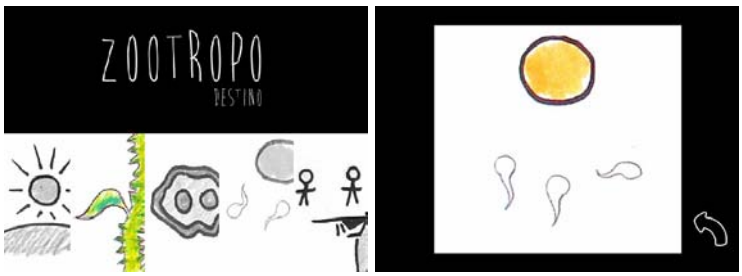


Fig. 5. El concepto seleccionado para la composición de este collage es “El Destino”.

Asimismo, se indagó la linealidad y no-linealidad en la narrativa, tal como se presenta en [15], anteponiendo nuevamente, el discurso del relato por encima de la herramienta.

5 Conclusiones y trabajos futuros

El flujo de la red ha estimulado el surgimiento de nuevas prácticas de comunicación mediadas informáticamente con diversos grados de interactividad y posibilidades de crear comunidades, globales y locales. Los artistas han incorporado estas prácticas a sus propuestas. Por tanto, esta experiencia propone una configuración diferente en el uso de las TIC, como una herramienta insoslayable en una disciplina particular, como son las Artes Combinadas, intentando estimular las prácticas sin “corromper” de alguna manera la creatividad de los estudiantes.

Con ello, se pretende explorar la creatividad y al mismo tiempo, lograr una integración de contenidos introductorios, para dar continuidad a los demás niveles curriculares.

Para el caso de estudio expuesto, los resultados alcanzados en la cohorte 2013, demuestran la aprehensión de los conceptos y la comprensión de los objetivos presentados a partir de las actividades diseñadas y las estrategias propuestas.

El ET1, si bien requiere una lectura comprensiva por parte de los estudiantes, esta actividad se considera imprescindible para la profundización y reflexión acerca de las relaciones de ciencia, arte y tecnologías.

Sin lugar a duda, la "invasión" de las TIC, favorecen en gran medida la facilidad de su uso, por lo que indagar en conceptos relacionados a la abstracción de las ideas y los problemas fue la dificultad que se presentó con mayor frecuencia en los grupos. Por tanto, se considera vital, una estimulación constante en la creatividad de los estudiantes y fortalecer la retroalimentación en las etapas proceso del creativo, en relación al ET2; lo cual significará una mejor producción tanto en dicho eje, como en el ET3.

Referencias

1 Resumen Descriptivo Licenciatura En Artes Combinadas. Universidad Nacional del Nordeste, http://www.artes.unne.edu.ar/documentos/Artes-Combinadas.pdf

2 Schiavo E.: Investigación científica y tecnológica en el campo de las TIC: ¿conocimientos técnicos, contextuales o transversales? Rev. Iberoam. Cienc. Tecnol. Soc. v.3 n.9 Ciudad Autónoma de Buenos Aires. (2007)

3 Blanco, J., Garcia, P. y Cherini, R.: Convergencias y divergencias en la noción de computación. Rev. Iberoam. Cienc. Tecnol. Soc. [online], vol.7, n.19, pp. 111-121. ISSN 1850-0013. (2012)

4 Martínez, A. García-Beltrán, Breve historia de la informática r. División de Informática Industrial ETSI Industriales – Universidad Politécnica de Madrid C/ José Gutiérrez Abascal, 2. 28006 – Madrid (España)

5 Manovich, L.: The Language of New Media. http://www.manovich.net/LNM/Manovich.pdf. (2001)

6 Kurzweil R.: The Age of Intelligent Machines "Chronology" http://www.calculemus.org/lect/si/dlalomzy/mchron.htm

7 Manovich L.: La vanguardia como software. Departamento de Artes Visuales (Universidad de California). (1999)

8 Warner Marien M.: *Photography: A Cultural History*. *ISBN*-10: 1856696669, *ISBN*-13: 978-1856696661, Edition: 2.(2010)

9 Computer History Museum, http://www.computerhistory.org/babbage/

10 Babbage, H.: Babbage's Calculating Engines. (New York). Edicion impresa Cambridge University Press. ISBN: 1108000967, 9781108000963. (2010)

11 Giudice, J.: Complejidad y dimensiones en los estudios sobre Babbage: la máquina analítica. Un análisis del fracaso cultural del primer proyecto de calculadora digital programable secuencialmente. Rev. Española de ciencia, tecnología y sociedad, y filosofia de la tecnología, ISSN 1139-3327, Nº 4, http://institucional.us.es/revistas/argumentos/4/art_1.pdf. (2010).

12 Eames, Charles, *A Computer Perspective: Background to the Computer Age*, Cambridge (Massachusetts), Harvard University Press. (1990).

13 García Rolando. Sistemas complejos. Conceptos, método y fundamentación epistemológica de la investigación interdisciplinaria, Barcelona, Gedisa, 200 pp. ISBN 94-9784-164-6 (2006)

14 Causa E., Romero Costas M., Rivero E., Bedoian D.: Proyecto *Biopus,* http://www.biopus.com.ar/

15 Alé, G; Sosa, F; Verrier, F.: La ruptura de la linealidad en el relato Vanguardias, Videoarte, Net Art. Facultad de Diseño y Comunicación. Universidad de Palermo. Buenos Aires, Argentina. ISSN 1668-5229. (2004)

Aplicación del aprendizaje basado en problemas y la tecnología informática a la enseñanza de programación en los primeros años de ingeniería

Ricardo Coppo[1], Javier Iparraguirre[1], Germán Feres[1], Gustavo Ursua[1], and Ana Cavallo[2]

[1] Universidad Tecnológica Nacional - Facultad Regional Bahía Blanca
[2] Instituto Superior Juan XXIII - Bahía Blanca
Mail de contacto: rcoppo@frbb.utn.edu.ar

Resumen La enseñanza de la programación de computadoras es una materia básica de las carreras relacionadas con las ciencias de la computación y de la ingenierías electrónica y de sistemas. Algunos estudios sugieren la necesidad de incrementar los factores motivacionales causados por la resolución exitosa de problemas relacionados con la carrera universitaria elegida, y que aumenten el entusiasmo, autoestima y perseverancia del alumno. Se evalúa la inserción en la dinámica del curso de programación tradicional un elemento didáctico, basado en una placa electrónica microcontroladora de licencia open source, a la que se le ha adicionado desarrollos de hardware propios realizados por la cátedra. Se trabajó con una metodología didáctica basada en la resolución de problemas (PBL) con el fin de observar su incidencia en la calidad del aprendizaje y la generación de motivación intrínseca por parte de los alumnos. Las primeras experiencias muestran un fuerte incremento en el desempeño de los alumnos y mejores resultados en los proyectos finales de cátedra.

Palabras clave: enseñanza de la programación, hardware didáctico, motivación.

1. Introducción

La mayoría de los educadores y docentes de programación coinciden en señalar que son pocos los estudiantes que afirman que aprender a programar una computadora es una tarea sencilla. El problema se agrava debido a que la mayoría de los cursos de programación se encuentran en los currículos de los primeros años de las carreras de

Ingeniería Electrónica, años que significan para la mayoría de los estudiantes transiciones importantes en sus hábitos de estudio, de vida, y comportamiento social [4,6].

La literatura especializada enfoca los avances de la enseñanza de la programación con el uso de tecnologías nóveles (proyecciones, multimedios, y otros) que pueden ser aplicadas a la didáctica del docente y en la organización de sus clases [7,8]. Menos atención ha sido dedicada a los factores cognitivos presentes en el proceso de aprendizaje. El estilo de aprendizaje y la motivación se destacan como los factores cognitivos que afectan en mayor medida el desempeño del alumno.

La motivación, que por definición es "la dirección y magnitud del comportamiento humano", tiene como factores indispensables la elección del curso de acción, la persistencia en el mismo y el esfuerzo que el alumno decidirá realizar [1,5].

Una vez decidido el curso de acción el factor emocional/afectivo tiene un rol preponderante. El alumno se relaciona no solo cognitivamente con su proyecto sino también en forma afectiva; característica que es fundamental para incrementar el tiempo y el esfuerzo que luego dedicará a la conclusión del mismo.

Tradicionalmente la motivación ha sido considerada bajo una perspectiva individualista. Sin embargo, como toda acción humana se realiza en un contexto, este ejerce una influencia indiscutible en todo el proceso de aprendizaje. Además, las exigencias de la sociedad moderna priorizan las actitudes colaborativas y la interacción entre pares [1].

En una primera evaluación, la mayoría de los docentes de informática, presienten que el aprendizaje de la programación en estos primeros años obedece más a una motivación extrínseca, como aprobar la materia con altas calificaciones o lucir como intelectual entre sus pares. Sin embargo, los alumnos que demuestran un verdadero interés personal en adquirir la capacidad de aprender a programar correctamente (motivación intrínseca) son los que luego logran los mejores resultados.

Mediante la aplicación de las metodologías de aprendizaje basado en problemas PBL *(Problem Based Learning)* [1,2] se intenta crear un contexto interactivo en el que al trabajar sobre la resolución de problemas individuales, se incrementa la autoestima y la coopera-

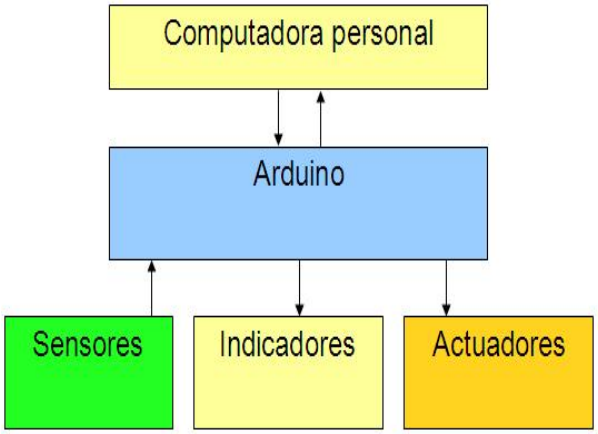


Figura 1. Diagrama en Bloques

ción entre los alumnos, integrando conocimientos adquiridos en otras áreas o materias.

Visto esto, este trabajo intenta explorar el uso de un sistema de hardware y software que introducen fuertes factores de motivación intrínseca en los alumnos de ingeniería electrónica, generándoles verdaderos deseos de poder desarrollar programas de software y palpar los resultados en equipos electrónicos reales.

Esto último además transfiere al estudiante un elemento de distinción entre sus pares (especialmente con los alumnos de otras carreras) que fortalece su autoestima y su decisión en la elección de la carrera elegida.

2. Materiales y metodología

2.1. Descripción de sistema

El sistema didáctico utilizado se divide en los siguientes subsistemas: un software de desarrollo que se ejecuta sobre una laptop o PC estándar y una placa de hardware Arduino interconectada con la computadora y diversos sensores o actuadores desarrollados por la cátedra que le permiten a la placa interactuar con el entorno o medio físico. (Figuras 1 y 2).

Figura 2. Sistema completo

2.2. Características del hardware

El elemento principal del hardware del sistema consta de una plataforma open-source denominada Arduino UNO [9], basada en un microcontrolador que dispone de 14 entradas/salidas digitales, 6 entradas analógicas, una interfase USB y, 32 KBytes de memoria para el programa.

La gran cantidad de entradas y salidas permite construir pequeños experimentos con sensores básicos de temperatura, LEDs, y displays de bajo consumo sin tener que recurrir a conexionados complejos. (Figura 3). La placa se monta en un soporte de acrílico para incentivar al alumno a preguntar sobre sus componentes y el funcionamiento del mismo. Las conexiones externas aceleran la configuración del sistema para diferentes ejercicios. (Figura 4).

El hardware open-source es aquel cuyo diseño y esquemas circuitales se publican, por ejemplo a través de la Internet, para que cualquier persona o empresa pueda fabricarlo o reproducirlo sin pagar licencias. Este modelo comercial se encuentra sustentado por el concepto del software open-source propulsado por los programadores del sistema operativo GNU/Linux.

La libre disponibilidad de la información de diseño permite a diversos proveedores de hardware la creación y venta de kits de compo-

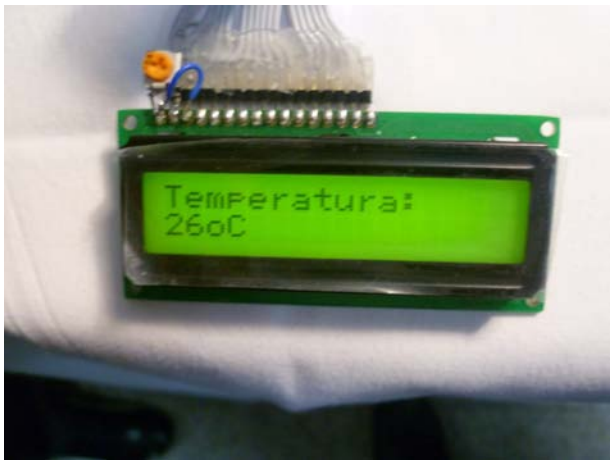


Figura 3. Control de temperatura

nentes para su armado, y posibilitan al alumno entusiasta construir su propia plataforma de experimentación a bajo costo.

Gracias a la característica open-source del hardware los docentes de la cátedra pudieron extender las facilidades básicas del sistema con desarrollos propios. La primera, una hilera de LEDS (1 sola dimensión espacial) permiten desarrollar aplicaciones como el "auto fantástico" o secuenciadores de luces como las que se instalan sobre patrullas policiales y de ambulancias. Una segunda extensión basado en una matriz de LEDS de dos dimensiones permite explorar problemas relacionados con ciclos y "recorridas" de matriz además de presentar conceptos iniciales sobre el funcionamiento de un display o monitor digital. (Figuras 5 y 6).

2.3. El ambiente de desarrollo

La programación de la placa se realiza por medio de un entorno de desarrollo integrado (IDE) de distribución gratuita que permite al usuario escribir aplicaciones en C que es uno de los objetivos principales del programa sintético de la materia. Una vez compiladas estas últimas, son transmitidas a la placa por medio del puerto USB a través del IDE. Es de destacar la importancia que el hardware reconoce esencialmente el mismo lenguaje de programación que los alumnos estudian en las clases teóricas.

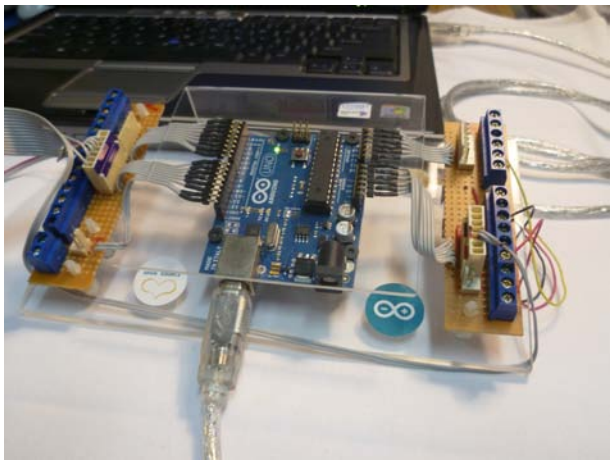


Figura 4. Placa y conexion a la PC

3. Aplicación educativa

En las primeras clases de la materia el alumno aprende sobre la arquitectura de las computadoras y los pasos necesarios para producir un programa ejecutable. Los primeros ejercicios se desarrollan con entrada/salida en modo terminal, es decir sobre la antigua pantalla del DOS histórico. Conceptos básicos como el sistema binario, la noción de algoritmo, y las primeras instrucciones de bifurcación y ciclo en el lenguaje C resultan abstractos y difíciles de entender para la mayoría de los estudiantes no motivados.

Un ejercicio clásico, realizado en la mayoría de las cátedras de programación para fijar estos conceptos, es la realización de un programa en el que se hace la conversión de números del sistema decimal al sistema binario (base 2) con salida a pantalla de línea de comando en modo terminal. El ejercicio, de nivel conceptual interesante para aprender a programar, no resulta motivador para el estudiante porque el alumno percibe su salida como algo antiguo o de poca utilidad.

Para realizar este mismo ejercicio sobre la placa de hardware solo bastan conectar algunos diodos emisores de luz (LED) a las salidas correspondientes y modificar ligeramente el programa originalmente destinado a la pantalla DOS para que opere con los LEDS.

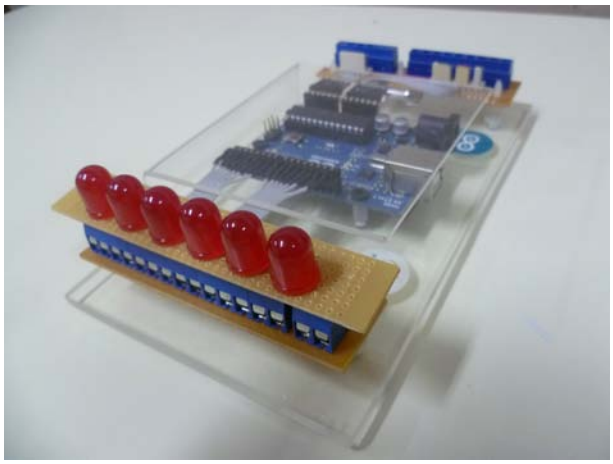


Figura 5. Hilera de LEDS - Problemas 1D

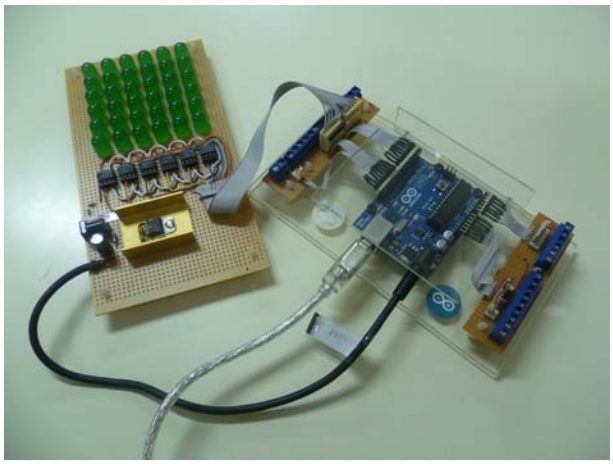


Figura 6. Matriz de LEDS - Problemas 2D

Frecuentemente el primer intento del programa no funciona correctamente (luces fuera de secuencia, lógica errónea, entre otras fallas), pero a diferencia del ejercicio equivalente en pantalla, el alumno se motiva más, quiere ver su código funcionando, muestra mayor tezón en la búsqueda de la solución, y se retira del laboratorio con mayor satisfacción personal por haber cumplido el objetivo.

Cabe señalar que la configuración presentada se adapta a ejercicios más complicados y a metodologías didácticas diferentes. Por

ejemplo, fácilmente la configuración presentada se puede adaptar a la enseñanza de autómatas regulares simulando sistemas de alarmas, control de semáforos, electrodomésticos inteligentes, por mencionar solo algunas ideas. Aunque estos temas se estudian en mayor profundidad en materias más avanzadas en la carrera, percibir como es su funcionamiento despierta mucho interés en el alumno que recién inicia y además le permite visualizar el tipo de sistema que podrá diseñar en un futuro cercano.

4. Resultados

El sistema propuesto se puso en marcha en forma experimental en la cátedra Informática I de la Universidad Tecnológica Nacional - Facultad Regional Bahía Blanca durante los ciclos lectivos 2011 y 2012. En el primer año se disponía de un solo equipo al que se le aplicó la técnica "open shop batch" en que cada par de alumnos disponía de 20 minutos para resolver un problema específico. La respuesta obtenida alentó la publicación de los primeros resultados y a iniciar acciones para la adquisición de mayor cantidad de equipos.

En el año 2012 se efectuaron más laboratorios con el nuevo hardware de soporte y con mayor cantidad de sensores y LEDS que permitieron plantear problemas de mayor complejidad. La aparición comercial a finales de dicho año de equipos similares de aún menor costo que el original produjo dos resultados sorprendentes. La primera es la donación por parte de los alumnos de 10 equipos para la cátedra, indicando que aunque no eran para su provecho directo, advertían su importancia para los futuros alumnos de la materia en años posteriores. El segundo fue la adquisición de mas de 30 equipos adicionales para uso personal de los alumnos a quienes se les despertó el interés por desarrollar sus propios proyectos.

También se observó una sensible mejora en los trabajos de fin de materia en los años 2011, año en que uno de los trabajos recibió el importante segundo premio en el concurso estudiantil de las jornadas JAIIO 40 - Jornadas Argentinas de Informática e Investigación Operativa desarrollado en la ciudad de La Plata [10]. La asistencia al congreso y su participación fue un factor emotivo y motivante para el asistente. Otros proyectos presentaron ese año incluyeron juegos

electrónicos combinando desarrollos incipientes en hardware y software y manejo de bases de datos simples entre otros.

En el grupo de los alumnos del año 2012 los proyectos finales de software presentan una complejidad que triplica, evaluada en cantidad de líneas de código, la producción de años anteriores.

5. Conclusiones

Las primeras experiencias con la metodología presentada han sido muy positivas. La participación y motivación de los alumnos al ver sus proyectos convertidos en componentes y diseños reales incentiva la creatividad y búsqueda de proyectos de mayor envergadura. El uso de un hardware comercialmente económico y con gran difusión en Internet permite al alumno despertar hábitos de investigación y búsqueda de problemas que (aunque resueltos) pueden ser implementados con bajo costo. Un proyecto exitoso deriva en otro y así sucesivamente; crece la autoestima y brinda elementos de valoración entre sus pares.

En nuestro diseño prevaleció la idea de experimentar con diversos sensores y medios de salida, en un esfuerzo para determinar experimentos motivadores y significativos para la enseñanza de la programación. En una versión final, más apta para cursos numerosos, se debería definir una configuración básica de sensores y construir todo el sistema físicamente en una sola unidad. Eventualmente se podría extender el sistema básico a través de módulos independientes para periféricos más complejos o de mayor potencia. El sistema propuesto debería ser acompañado por dos documentos con diferentes tipos de ejercicios propuestos, uno para el alumno y otro para el profesor.

Referencias

1. Dörnyei, Z., *Teaching and Researching Motivation*, Pearson Education, Malaysia, 2001.
2. O'Kelly J., Gibson J.P., *Robocode & Problem Based Learning: A non-prescriptive approach to teaching programming*, In ITICSE '06: Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education, 2006.
3. Lin H.T., Kuo T.H., *Teaching programming technique with edutainment robot construction*, IEEE 2nd International Conference on Education Technology and Computer (ICETC), Shanghai, June 2010.

4. Huet I., Pacheco O.R., Tavares J., Weir G., *New Challenges in Teaching Introductory Programming Courses: a Case Study*, 34th ASEE/IEEE Frontiers in Education Conference, Savannah, Georgia, October, 2004.
5. Jenkins T., *Teaching Programming - A Journey from Teacher to Motivator*, 2nd Annual LTSN-ICS Conference, London, 2001.
6. Jenkins T., *On the Difficulty of Learning to Program*, 3rd Annual LTSN-ICS Conference, Loughborough University, 2002.
7. Matsumura K., Daisuke S., Aiguo HE, *A C Language Programming Education Support System base on Software Visualization*, IEEE Joint Conferences on Pervasive Computing (JCPC), Tamsui, Taipei, December 2009.
8. Shyu Y.H., Chen P.W., *PLL: A Programming Languages Lab System*, 21st International Conference on Distributed Computing Systems Workshops (ICDCSW01), Mesa, Arizona, USA, 2001.
9. Arduino Team, *Arduino Home Page*, http://www.arduino.cc, visitado: Marzo 2013.
10. Diaz, A., Coppo R.J., *Calculadora geoespacial*, Jornadas Argentinas de Informática e Investigación Operativa, JAIIO 40, Segundo Premio Concurso Estudiantil, La Plata, Argentina, 2012.

Construcción de Clasificadores Automáticos de Habilidades de E-Tutores de Aprendizaje Colaborativo

Pablo Santana Mansilla [(1,2)], Rosanna Costaguta [(2)], Daniela Missio[(2)]

(1) CONICET, Comisión Nacional de Investigaciones Científicas y Técnicas
(2) Instituto de Investigación en Informática y Sistemas de Información (IIISI), Departamento de Informática, Facultad de Ciencias Exactas y Tecnologías (FCEyT),
Universidad Nacional de Santiago del Estero (UNSE),
Avda. Belgrano (S) 1912, Santiago del Estero, 4200, Argentina
{psantana, rosanna, dmissio}@unse.edu.ar

Resumen. El Aprendizaje Colaborativo Soportado por Computadora (ACSC) permite a los estudiantes aprender por medio de la interacción con personas localizadas en marcos temporales y espaciales diferentes. Disponer de software especializado que apoye el aprendizaje en grupo, no garantiza que los estudiantes colaboren, porque no es la tecnología utilizada lo que soporta la colaboración sino la forma en que la utilizan los e-tutores (docentes) para coordinar la interacción grupal. Teniendo en cuenta que el e-tutor es clave para el éxito del ACSC, y que poco se sabe sobre como los docentes intervienen en las actividades de aprendizaje colaborativo de los estudiantes, se creó en primera instancia un esquema de clasificación de habilidades de e-tutores de ACSC. Considerando que el análisis manual de las interacciones registradas en entornos de ACSC requiere considerable tiempo y esfuerzo, en segunda instancia se desarrollaron y evaluaron métodos automáticos de análisis de interacciones y reconocimiento de habilidades.

Palabras Clave: Aprendizaje Colaborativo Soportado por Computadora, e-tutor, clasificación de habilidades de e-tutor, reconocimiento automático de habilidades.

1 Introducción

Con el rápido desarrollo de la sociedad basada en conocimiento ha crecido la importancia que se le da a la creación colaborativa de conocimientos. En este contexto, el ACSC se ha vuelto una nueva forma de educación donde los estudiantes aprenden por medio de la interacción con personas que pueden estar localizadas en marcos temporales y espaciales diferentes [1]. Por otro parte, las investigaciones realizadas demuestran que los sistemas de ACSC brindan un entorno apropiado para el desarrollo de habilidades tales como solución de problemas, pensamiento crítico, establecimiento de metas, interpretación, y análisis [1, 2]. Sin embargo, disponer de herramientas de software que soporten el aprendizaje en grupo no garantiza que los estudiantes colaboren, porque no son las características de la tecnología utilizada sino la forma en que se la utiliza lo que soporta la colaboración [3, 4, 5].

Según [6] y [7] tanto la calidad como la profundidad del aprendizaje dependen de la forma en que la colaboración es coordinada pero, la mayoría de los e-tutores no están familiarizados con las técnicas que se requieren en ACSC, ni se conocen con precisión las habilidades que deberían poseer para desempeñarse adecuadamente [5, 7, 8]. Si además se tiene en cuenta que las iniciativas de ACSC tienen pocas chances de éxito sin docentes con habilidades para sacarle provecho a las herramientas tecnológicas disponibles [5, 7], es evidente la necesidad de plantear mecanismos que les permitan a los e-tutores adquirir las habilidades necesarias para desempeñarse adecuadamente en entornos de ACSC.

De acuerdo con [9] una manera de propiciar la adquisición de habilidades consiste en desarrollar sistemas de software que entrenen a los e-tutores. No obstante, tanto la falta de conocimiento sobre las habilidades requeridas en un e-tutor colaborativo, como los efectos negativos sobre la comunicación que presentan las técnicas de análisis de interacciones basadas en interfaces estructuradas o semi estructuras [10, 11], constituyen los principales obstáculos a vencer para desarrollar aplicaciones que permitan a los e-tutores adquirir las habilidades que manifiestan con deficiencia.

En el marco del proyecto de investigación "Sistemas de información web basados en agentes para promover el Aprendizaje Colaborativo Soportado por Computadoras" (CICyT-UNSE Código 23/C097), se dieron varios pasos tendientes a propiciar el futuro desarrollo de sistemas informatizados que entrenen a los e-tutores de ACSC. Primero, se definió un esquema de clasificación de las habilidades que debería poseer un e-tutor de ACSC. Luego, de forma complementaria a la taxonomía de habilidades creada, se aplicaron técnicas de minería de textos para construir clasificadores con la capacidad de identificar automáticamente las habilidades manifestadas por e-tutores al interactuar con sus alumnos. El trabajo realizado se documenta en el presente artículo que se estructura como sigue. La próxima sección describe brevemente el esquema de clasificación de las habilidades que deberían poseer los e-tutores de ACSC. La sección 3 presenta las dos maneras en que puede realizarse el análisis de las interacciones en entornos de ACSC (manual y automático). La sección 4 describe la aplicación de la técnica de análisis de contenido. La sección 5 comenta las adaptaciones realizadas a GATE[1] (la herramienta de minería de textos empleada) a fin de procesar textos en español. La sección 6 especifica los pasos metodológicos ejecutados para construir los clasificadores, la experimentación realizada, y los resultados obtenidos con diferentes algoritmos de clasificación. Por último, en la sección 7 se enuncian algunas conclusiones.

2 Clasificación de Habilidades de E-Tutores

A pesar de que las habilidades propias de un e-tutor de ACSC no han sido descriptas detalladamente, entre los investigadores que han estudiado esta temática se pueden identificar dos enfoques. Por un lado, se proporcionan guías y recomendaciones para e-tutores de ACSC pero sin brindar mayores precisiones. Por el otro lado, se analiza el rol de los e-tutores de ACSC tomando como base el esquema de organización por roles, heredado de la literatura sobre habilidades y competencias para docentes de

[1] http://gate.ac.uk/

e-learning que no recurren al aprendizaje colaborativo como estrategia pedagógica sino que se basan en el modelo de interacción docente-estudiante. Como parte de la creación del esquema de clasificación de habilidades de docentes de ACSC en [12] se integraron los dos enfoques citados, teniendo en cuenta además los planteos pertenecientes al área del *e-learning* no colaborativo. Con el propósito de facilitar la comprensión de la labor de los docentes de ACSC se organizó la taxonomía en tres niveles (Figura 1a). El primer nivel responde a las habilidades vinculadas con los roles que desempeñan los e-tutores al coordinar el aprendizaje: Administrativo, Pedagógico, Social, Técnico, Comunicación y Evaluación (Figura 1b). El segundo nivel considera las sub habilidades relacionados con cada habilidad, y finalmente, el tercer nivel especifica atributos para cada sub habilidad.

Si bien en la Figura 1b se muestran seis categorías de habilidad, se trabajó sólo con las habilidades sociales para evaluar la viabilidad del uso de la minería de textos en el reconocimiento automático de habilidades.

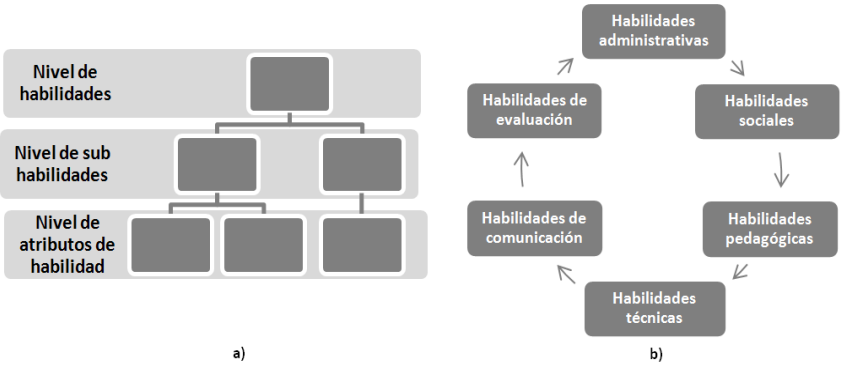


Fig. 1. Niveles y categorías de la clasificación de habilidades.

3 Análisis de Interacciones en ACSC

Tomando como base la clasificación de habilidades descripta en [12] se podrían analizar las contribuciones docentes en un ambiente ACSC a fin de identificar las habilidades manifestadas. Pero, tal proceso de análisis no es sencillo ya sea que se realice de forma manual o automática. Por una parte, cuando se tiene una cantidad considerable de interacciones el análisis manual es prácticamente inviable debido al tiempo y el esfuerzo que demanda [13]. Por otra parte, en el análisis automático de la información registrada por un sistema de ACSC se ha recurrido tanto a interfaces estructuradas o semi estructuradas como a texto libre para modelar las interacciones. Las interfaces estructuradas y semi estructuradas (oraciones de apertura, contribuciones dirigidas por menú, y contribuciones basadas en diagramas) simplifican el proceso de análisis al no tener que usar técnicas de procesamiento de lenguaje natural [10, 11, 14]. Sin embargo, este tipo de interfaces socava el proceso de colaboración porque restringe las posibilidades de interacción, provoca que la

comunicación sea lenta, y crea stress relacional [3, 10, 11]. En relación al texto libre, al permitir que la comunicación se realice sin restricciones ya no es posible usar la interfaz para hacer inferencias sobre el proceso de interacción, y por consiguiente, es necesario recurrir a técnicas de procesamiento de lenguaje natural. Hasta el momento, este tipo de tecnología se usó de manera limitada en tareas como la clasificación de los tópicos de una conversación y la caracterización de patrones de discusión [14].

Debido a las deficiencias asociadas con el análisis manual y con el análisis automático de interacciones estructuradas o semi estructuradas, surgió la necesidad de contar con una técnica capaz de reconocer las habilidades que poseen los e-tutores sin afectar negativamente la dinámica de trabajo de los estudiantes o docentes, ni representar un alto costo o sobrecarga. En este contexto la minería de textos (en particular sus técnicas de clasificación) se vislumbró como una alternativa para cumplir con las condiciones mencionadas, dada su capacidad para manejar la incertidumbre, borrosidad, diversidad de estructuras, y gran cantidad de palabras que caracterizan al lenguaje natural [15].

4 Aplicación de Análisis de Contenido

Para obtener el conjunto de datos (entrenamiento y prueba) que se usó para aplicar minería de textos, se realizaron varias experiencias de ACSC en el entorno de aprendizaje Moodle[2] con docentes y estudiantes de la FCEyT de la UNSE. En pos de incrementar la cantidad de datos usados en la construcción de los clasificadores se solicitó la colaboración de investigadores que participaron del proyecto Tactics [16], quienes brindaron acceso a las sesiones de chat y comunicaciones vía foro de las experiencias de ACSC en las que estuvieron involucrados. Sin embargo, las interacciones por sí mismas no son suficientes para construir los clasificadores porque también se necesitan conocer los atributos de habilidad manifestados por los e-tutores. Por consiguiente, sobre el conjunto de interacciones se aplicó análisis de contenido (siguiendo el planteamiento metodológico de Krippendorff [17]). Así, expertos humanos (un psicopedagogo y un especialista en ACSC) hicieron corresponder un atributo de habilidad social de la clasificación comentada en la sección 2 a cada oración que formaba parte de los mensajes o contribuciones de los docentes. Cabe aclarar que los expertos no tuvieron que catalogar todos los mensajes publicados por los e-tutores sino solamente aquellos pre seleccionados mediante la técnica de muestreo de relevancia [17] por considerar que contenían al menos una oración donde se manifestaba un atributo de habilidad social. El nivel de acuerdo entre los expertos humanos (para las 891 oraciones resultantes) fue de $\alpha = 0.938$. Si bien este valor es superior al 0.80 recomendado por Krippendorff para aceptar los resultados de un estudio de análisis contenido como fiables, no representaba un nivel de acuerdo perfecto ($\alpha = 1$) sino que indicaba que existían oraciones a las que los expertos asignaron atributos de habilidad diferentes. Para resolver estas discrepancias se realizó un segundo proceso de análisis donde los expertos unificaron criterios y

[2] https://moodle.org/

acordaron el atributo de habilidad que le correspondía a cada oración sobre la que existía desacuerdo.

5 Adaptaciones Realizadas a GATE

Teniendo en cuenta que en cada mensaje o contribución de los e-tutores se pueden manifestar múltiples habilidades, como paso previo a la construcción de los clasificadores, fue necesario dividir en oraciones a los mensajes. La descomposición de los mensajes de los e-tutores en oraciones tiene influencia directa en los resultados que se obtienen con los clasificadores de habilidades ya que una identificación errónea de oraciones llevaría a que las habilidades no se reconozcan apropiadamente. Si bien GATE cuenta con recursos de procesamiento para identificar oraciones en un documento, los mismos fueron creados para ser aplicados a textos en inglés y por consiguiente fue preciso adaptarlos para que trabajen correctamente al procesar textos en español.

GATE ofrece dos recursos de procesamiento para la segmentación de textos en oraciones: *RegEx Sentence Splitter* y *ANNIE Sentence Splitter* [18]. El primero de estos tuvo que ser descartado ya que no se pueden reconocer correctamente oraciones cuando se tiene abreviaturas seguidas de signos de puntuación. En [19] se brindan detalles sobre las reglas que se crearon de modo tal que mediante *ANNIE Sentence Splitter* se puedan identificar oraciones en textos en español. Las *stop words* son palabras de uso frecuente en un lenguaje (artículos, preposiciones, conjunciones, pronombres, etc.) que pueden ser descartadas porque tienen muy poco contenido de información para distinguir entre categorías y una capacidad predictiva escasa [15, 20]. GATE no cuenta con un recurso dedicado exclusivamente a la supresión de *stop words* por lo cual en [19] también se describe como implementar esta operación de pre procesamiento mediante *JAPE Transducer* y *ANNIE Gazetteer* [18].

6 Construcción de los Clasificadores

La minería de textos o *text mining* se refiere al proceso de extracción de patrones interesantes y no triviales o conocimiento desde documentos de texto [15, 21]. La minería de textos intenta revelar la información oculta por medio de métodos que son capaces de tratar con la vaguedad, incerteza, borrosidad, y gran cantidad de palabras y estructuras que caracterizan al lenguaje natural [15]. Si bien la minería de textos utiliza técnicas de minería de datos, en el proceso de descubrimiento de conocimiento la minería de textos parte de datos textuales no estructurados mientras que la minería de datos se aplica sobre bases de datos estructuradas [22]. Las técnicas de clasificación de la minería de textos, que consisten en asignar objetos a categorías predefinidas, se adecuan naturalmente al problema de identificar las habilidades manifestadas por los profesores de ACSC, porque la intención es vincular cada contribución de los e-tutores con una o más habilidades de la clasificación propuesta en [12].

Realizadas las adaptaciones a GATE que se comentaron en la sección 5, el paso siguiente en la investigación consistió en construir los clasificadores siguiendo la

metodología CRISP-DM [23]. Si bien CRISP-DM fue propuesta para guiar el desarrollo de proyectos de minería de datos, puede usarse para abordar problemas de minería de textos en la medida que los datos textuales se transformen en un formato estructurado o semi estructurado [24].

La primera tarea de preparación de los registros de las sesiones de ACSC para la minería de textos consistió en convertirlos a texto plano. Las sesiones de ACSC estaban almacenadas en 4 formatos de archivo (htm, rtf, doc, y pdf) y para evitar problemas de compatibilidad en la representación de caracteres los mensajes muestreados se guardaron en archivos de texto plano con un archivo por mensaje. A continuación se aplicaron las siguientes tareas de limpieza de datos: corrección de errores de ortografía, reemplazo de emoticones por caracteres Unicode equivalentes, supresión de archivos duplicados, y eliminación de archivos que contenían solo nombres de lugares o de personas por cuanto no brindan información relevante para la clasificación.

Los sistemas de minería de textos no aplican sus algoritmos de descubrimiento de conocimiento a colecciones de documentos no estructurados por lo tanto, es necesario recurrir a operaciones de pre procesamiento o de preparación de datos de manera de transformar documentos de texto no estructurados en un formato intermedio estructurado más explícitamente [20, 22]. Las operaciones de pre procesamiento se centran en la identificación, extracción, refinamiento y adición de características representativas de los documentos en lenguaje natural de modo tal que, las características más representativas sean usadas para el *text mining* mientras que las restantes descartadas [22]. En el presente trabajo de investigación se recurrió a las operaciones de pre procesamiento: identificación de token, identificación de oraciones, normalización (*inflectional stemming* y lematización), *Part-of-Speech Tagging* (POST), y supresión de *stop words* [15, 20, 22]. Estas operaciones no se aplicaron de manera aislada sino que se combinaron para determinar si tenían influencia en la efectividad de los algoritmos de clasificación. La Tabla 1 muestra las combinaciones posibles de las cinco operaciones de pre procesamiento utilizadas.

Tabla 1. Combinaciones de operaciones de pre procesamiento

Combinaciones	Identificación de Token	Identificación de oraciones	Eliminación de *stop words*	Lematización	*Stemming to a root*	POST
Pre procesamiento 1	x	x				
Pre procesamiento 2	x	x	x			
Pre procesamiento 3 A	x	x		x		
Pre procesamiento 3B	x	x			x	
Pre procesamiento 4	x	x				x
Pre procesamiento 5	x	x	x			x
Pre procesamiento 6 A	x	x	x	x		
Pre procesamiento 6 B	x	x	x		x	
Pre procesamiento 7 A	x	x		x		x
Pre procesamiento 7 B	x	x			x	x
Pre procesamiento 8 A	x	x	x	x		x
Pre procesamiento 8 B	x	x	x		x	x

Cuando un clasificador se construye mediante minería de textos se necesita que un conjunto de documentos manualmente clasificados por expertos del dominio, sea divido en dos subconjuntos, uno de entrenamiento y otro de prueba. El clasificador se construye mediante un proceso inductivo, donde observando las características del subconjunto de entrenamiento se infieren las condiciones que documentos previamente no examinados deberían cumplir para ser clasificados bajo una u otra categoría [25]. Luego, la comparación de las decisiones de categorización realizadas por el clasificador con las efectuadas por expertos humanos (sobre el subconjunto de prueba) permite evaluar la efectividad de la clasificación automática. En el marco de la presente investigación, para evaluar la efectividad de los clasificadores en el reconocimiento de habilidades, se calcularon las métricas de precisión y recall con validación cruzada 10-fold [25]. La precisión indica el porcentaje de oraciones clasificadas correctamente entre todas las oraciones a las que se les asignó un atributo de habilidad, independientemente de si el clasificador no asignó atributos de habilidad a oraciones que debieron ser clasificadas [18, 22]. Por su parte, el recall señala el porcentaje de oraciones clasificadas correctamente entre todas las oraciones a las que se les debía asignar un atributo de habilidad, independientemente de cuantas clasificaciones erróneas se realizaron [18, 22].

Para la construcción de los clasificadores se utilizaron los algoritmos KNN, SVM, PAUM, Naïve Bayes y C4.5 [20, 22, 25, 26]. Un análisis detallado del desempeño de cada uno de los cinco algoritmos mencionados se brinda en [27], pero cabe resaltar que los clasificadores con los niveles más altos de precisión se obtuvieron con los algoritmos SVM y PAUM. En el caso del algoritmo SVM se decidió recurrir a una variante conocida como SVM con márgenes desiguales [28] puesto que para todos los atributos de habilidad social la cantidad de ejemplos de entrenamiento positivos era pequeña en relación a la cantidad de ejemplos negativos. Se probaron solo 4 valores del parámetro τ (0.3, 0.4, 0.5 y 0.6) puesto que fuera de los mismos la precisión o el recall asumen valores muy bajos. Los experimentos realizados combinando los 4 valores de τ con las diversas operaciones de pre procesamiento (cuyos resultados se resumen en Figura 2) indican que la preparación de datos tiene una influencia variable sobre la efectividad de la identificación de habilidades. Mientras que la eliminación *stop words* y POST tienen un efecto casi imperceptible sobre la precisión (Figura 2a) y el recall (Figura 2b), la normalización puede tanto ocasionar una disminución como un incremento de la precisión dependiendo de si aplica *inflectional stemming* o lematización. Por su parte, el algoritmo PAUM utiliza un margen positivo (τ_{+1}) y un margen negativo (τ_{-1}) para hacer frente a problemas de clasificación donde los ejemplos positivos son escasos en relación a los ejemplos negativos. La Figura 3 muestra las medidas de efectividad de solo cuatro combinaciones de valores de los márgenes puesto que son los que permitieron: obtener el valor más alto de precisión (τ_{+1}=-1 y τ_{-1}=-1.5), llegar al valor más alto de recall (τ_{+1}=10 y τ_{-1}=0.5), y tener la menor diferencia entre recall y precisión (τ_{+1}=5 y τ_{-1}= 0 al igual que τ_{+1}=10 y τ_{-1}=-0.5). Se pudo comprobar que las operaciones de pre procesamiento no incrementan la precisión sino que tienden a disminuirla (Figura 3a), mientras que el recall permanece relativamente estable (Figura 3b). En la Figura 3 también se aprecia que para valores positivos de τ_{+1} no existen diferencias significativas entre las cifras de precisión y recall, pero para valores negativos de τ_{+1} la precisión supera ampliamente al recall.

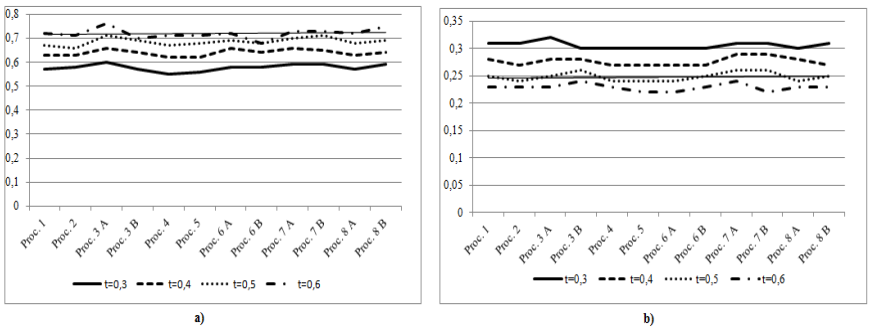


Fig. 2. Precisión y recall micro promediados para los clasificadores construidos con SVM

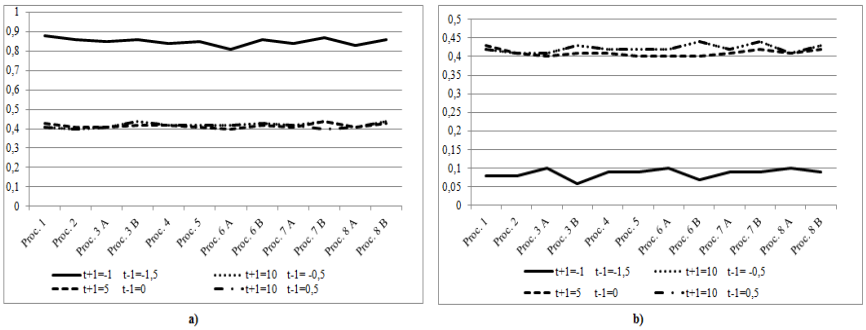


Fig. 3. Precisión y recall micro promediados para los clasificadores construidos con PAUM

7 Conclusiones

Durante la creación del esquema de clasificación de habilidades de e-tutores de ACSC quedó en evidencia que las distintas facetas del trabajo de un e-tutor (social, administrativa, pedagógica, técnica, comunicacional, y de evaluación) están vinculadas y es difícil separarlas. La taxonomía de habilidades se planteó con la intención de mejorar la comprensión de la tarea de los e-tutores pero, de ninguna manera puede considerarse definitiva sino que, por ser el fruto de un trabajo interdisciplinario con la pedagogía, se piensa que la clasificación podrá ser tomada como punto de partida por los profesionales del área interesados en realizar futuras investigaciones que permitan su validación científica. En este sentido, los resultados del análisis de contenido son auspiciosos porque en las experiencias de ACSC analizadas los expertos humanos reconocieron el 75% de los atributos de habilidad pertenecientes a la categoría social. Esto significa que en la práctica realmente se manifiestan los comportamientos y las conductas contempladas en la taxonomía de habilidades elaborada.

Teniendo en cuenta que los clasificadores automáticos construidos con los algoritmos PAUM y SVM permiten identificar habilidades en los mensajes de los

e-tutores con niveles de precisión superiores al 0.7, se puede afirmar que la minería de textos es una alternativa viable para conocer el desempeño de los e-tutores de ACSC sin demandar el tiempo y el esfuerzo requeridos para un análisis manual de las interacciones. A pesar de que se necesita seguir trabajando para mejorar el recall, los valores de precisión logrados son más que destacables considerando que, la cantidad de ejemplos utilizada para construir los clasificadores es muy inferior a la cantidad empleada en otros trabajos de investigación donde se utiliza minería de textos para tareas de clasificación. Así por ejemplo, en [26] con la colección Mod-Apte (una muestra de Reuters-21578 con 12902 documentos) se reportaron valores de precisión de 0.75, mientras que en la presente investigación se lograron niveles de precisión mayores al 0.7 con apenas 891 oraciones. Por otro lado, la experimentación realizada permite afirmar que la influencia de las operaciones de pre procesamiento sobre la efectividad del reconocimiento de habilidades depende del algoritmo de clasificación. Para obtener resultados verdaderamente concluyentes, se cree conveniente incrementar el número de interacciones, y también ampliar las categorías a reconocer ya que en esta investigación el reconocimiento se restringió a los atributos de la habilidad social.

Por último, los resultados obtenidos son alentadores sobre el uso de la minería de textos como parte de un fututo sistema software de entrenamiento personalizado de habilidades para e-tutores de ACSC. Una aplicación de este tipo tendría que analizar las interacciones grupales de modo de identificar conflictos que requieran la intervención de los e-tutores para resolverse. En estos casos, el sistema se encargaría de sugerir a los e-tutores las acciones a llevar a cabo para mejorar el aprendizaje grupal y simultáneamente practicar las habilidades que no hayan manifestado adecuadamente.

Referencias

1. Suh, H., Lee,S.: Collaborative Learning Agent for Promoting Group Interaction. ETRI Journal, 28(4), 461--474 (2006)
2. Day, T.W.: Online Collaborative Learning and Leadership Development. En: Rogers, P., Berg, G., Boettcher, J., Howard, C., Justice, L., Schenk, K. (eds) Encyclopedia of Distance Learning, Second Edition, pp. 1488--1492. Information Science Reference (2009)
3. Olivares, O. J.: Collaborative vs. Cooperative Learning: The Instructor's Role in Computer Supported Collaborative Learning. En: Orvis, K. L., Lassiter, A. L.R (eds) Computer-Supported Collaborative Learning: Best Practices and Principles for Instructors, pp. 20--39. Information Science Publishing, USA (2007)
4. Onrubia, J., Engel, A.: The role of teacher assistance on the effects of a macro-script in collaborative writing tasks. International Journal of Computer-Supported Collaborative Learning, 7(1), 161--186 (2012)
5. Orvis, K.L., Lassiter, A.L.R.: Computer-Supported Collaborative Learning: The Role of the Instructor. En: Ferris, S., Godar, S.H. (eds) Teaching and learning with virtual teams, pp. 158--179. Information Science Publishing (2006)
6. Borges, M.A.F., Baranauskas, M.C.C.: CollabSS: a Tool to Help the Facilitator in Promoting Collaboration among Learners. Educational Technology & Society, 6 (1), 64--69 (2003)
7. Kukulska-Hulme, A.: Do Online Collaborative Groups Need Leaders? En: Roberts, T.S. (ed) Online Collaborative Learning: Theory and Practice, pp. 262--280. Information Science Publishing (2004)

8. Greiffenhagen, C.: Making rounds: The routine work of the teacher during collaborative learning with computers. International Journal of Computer-Supported Collaborative Learning, 7(1), 11--42 (2012)
9. Barker P.: Skill set for online teaching. Word Conference on educational, Multimedia, Hipermedia, & Telecomunicatios (2002)
10. Chen, W.: Supporting teachers' intervention in collaborative Knowledge building. Journal of Network and Computer Applications, 29, 200--215 (2006)
11. Soller, A., Martínez, M. A., Jermann, P., Muehlenbrock, M.: From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. International Journal of Artificial Intelligence in Education, 15 (4), 261--290 (2005)
12. Santana Mansilla, P., Costaguta, R., Missio, D.: Habilidades de E-tutores en Grupos Colaborativos. En: Peñaranda, N., Zazarini, S., Bejarano, I. F. (eds) Experiencias Innovadoras en Investigación Aplicada, pp. 687--704. Ediciones DASS-UCSE, Jujuy (2012)
13. Rosé, C., Wang, Y., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning, 3 (3), 237--271 (2008)
14. Tchounikine, P., Rummel, N., McLaren, B.M.: Computer Supported Collaborative Learning and Intelligent Tutoring Systems. En: Nkambou, R., Bourdeau, J., Mizoguchi, R. (eds) Advances in Intelligent Tutoring Systems. SCI, vol. 308, pp. 447--463. Springer, Heidelberg (2010)
15. Hotho, A., Nürnberger, A., Paaß, G.: Brief survey of text mining. LDV Forum GLDV Journal for Computational Linguistics and Language Technology, 20 (1), 19--62 (2005)
16. Juárez Pacheco, M.: Recomendaciones para el uso académico de herramientas web gratuitas. Revista Mexicana de Investigación Educativa, 10 (25), pp. 577--584 (2005)
17. Krippendorff, K.: Content analysis: an introduction to its methodology, Second Edition. SAGE Publications, USA (2004)
18. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., et al.: Developing Language Processing Components with GATE Version 7 (a User Guide). Departamento de Ciencias de la Computación, Universidad de Sheffield, Reino Unido (2013)
19. Santana Mansilla P., Costaguta R., y Missio D.: Uso de la Arquitectura GATE para la Identificación de Sentencias en Textos en Español. Primer Congreso Argentino de la Interacción-Persona Computador@, Telecomunicaciones, Informática e Información Científica (2012)
20. Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J.: Text Mining Predictive Methods for Analyzing Unstructured Information. Springer, USA (2005)
21. García Adeva, J. J., Calvo, R. A.: Mining Text with Pimiento. IEEE Internet Computing, 10 (4), 27--35 (2006)
22. Feldman, R., Sanger, J.: The text mining handbook. Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2007)
23. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc, USA (2001)
24. Magalhães, S. E.: Descoberta de Conhecimento com o uso de Text Mining: Cruzando o Abismo de Moore. No publicado, Tesis de Master, Universidad Católica de Brasil (2002)
25. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1--47, (2002)
26. Li, Y., Zaragoza, H., Herbrich, R., Shawe-Taylor, J., Kandola, J.: The Perceptron Algorithm with Uneven Margins. En: Nineteenth International Conference on Machine Learning, pp. 379--386. Morgan Kaufmann Publishers Inc., San Francisco (2002)
27. Santana Mansilla P., Costaguta R., y Missio D.: Reconocimiento de Habilidades de Docentes en Aprendizaje Colaborativo Soportado por Computadora Mediante Minería de Textos. 42 Jornadas Argentinas de Informática e Investigación Operativa (JAIIO) (2013)
28. Li, Y., Shawe-Taylor, J.: The SVM with uneven margins and Chinese document categorization. En: 17th Pacific Asia Conference on Language Information and Computation (PACLIC17), pp. 216- 227. Colips Publications, Singapur (2003)

Manifestación de habilidades de colaboración en grupos de aprendizaje síncronos y asíncronos

Diego Yanacon Atia y Costaguta Rosanna

Instituto de Investigación en Informática y Sistemas de Información (IIISI)
Departamento de Informática, Facultad de Ciencias Exactas y Tecnologías (FCEyT),
Universidad Nacional de Santiago del Estero (UNSE),
Avda. Belgrano (S) 1912, Santiago del Estero, 4200, Argentina
diegopunk27@hotmail.com, rosanna@unse.edu.ar

Resumen. El uso de computadoras en el dominio del Aprendizaje Colaborativo permitió definir nuevos escenarios de enseñanza y de aprendizaje, dando origen así a los sistemas de Aprendizaje Colaborativo Soportado por Computadoras. Estos sistemas se centran en la dinámica de grupo para facilitar el aprendizaje. Sin embargo, constituir un grupo no es suficiente para que los estudiantes colaboren y realicen con éxito las tareas encomendadas. Una colaboración efectiva implica determinados comportamientos, los cuales están condicionados entre otros factores, por las habilidades colaborativas que los estudiantes sean capaces de manifestar durante la dinámica de trabajo. Esta investigación estudió la manera en que los estudiantes manifiestan sus habilidades de colaboración cuando operan medios sincrónicos y asincrónicos. Por medio de una herramienta software de comunicación y colaboración, especialmente desarrollada, que modela interacciones semiestructuradas utilizando oraciones de apertura, fue posible realizar diferentes sesiones de colaboración con estudiantes universitarios. El análisis de las interacciones registradas en estas sesiones permite afirmar que los estudiantes manifiestan sus habilidades de colaboración con determinadas diferencias cuando trabajan de manera síncrona y asíncrona.

Palabras clave: Aprendizaje Colaborativo Soportado por Computadora, habilidades de colaboración, comunicación síncrona, comunicación asíncrona, grupos de estudiantes

1 Introducción

Cuando dos o más personas aprenden algo juntas, o al menos lo intentan, Dillenbourg [1] establece que se produce una situación de aprendizaje colaborativo (AC). El AC se refiere a una situación en la que se espera que sucedan ciertas interacciones entre personas, las cuales promuevan mecanismos de aprendizaje, sin garantía de que esas interacciones esperadas ocurran. En particular, cuando esas interacciones se materializan a través del uso de computadoras, entonces se habla de Aprendizaje Colaborativo Soportado por Computadoras (ACSC).

El ACSC es un campo de investigación emergente, que se enfoca en cómo el AC junto con el soporte tecnológico adecuado, pueden mejorar la interacción y el trabajo

en grupo; y además, cómo la colaboración a través de redes de información, facilita el intercambio y la distribución de conocimientos y experiencias entre los estudiantes que trabajan en grupos [2]. Las tecnologías *groupware* que dichos entornos utilizan, permiten a los docentes usar herramientas de comunicación y coordinación para actividades tales como la preparación de sus clases, la programación de actividades, el envío de notas, la comunicación de ideas, el seguimiento y evaluación del aprendizaje. Del mismo modo, los estudiantes las usan principalmente para comunicarse con el objeto de realizar las tareas asignadas, pueden colaborar en la complementación de información pertinente en wikis, generar discusiones sobre un tema determinado en foros, entre otras actividades posibles.

En un ambiente de ACSC los estudiantes pueden interactuar aprovechando sus conocimientos y habilidades, sin embargo esto no implica que ellos colaboren de manera eficaz, ni tampoco que logren obtener los conocimientos y adquirir las destrezas que el docente espera. Esto se debe a que no siempre un estudiante tiene las habilidades de colaboración desarrolladas, resultando decisiva en muchos casos, la formación y orientación que reciben de parte del docente o tutor, tanto de forma previa como durante el proceso colaborativo [3, 4].

Existen diferentes estudios que confirman que la colaboración entre personas estimula el aprendizaje, incentiva la comunicación y acrecienta la motivación personal al aprender colaborativamente [5]. Sin embrago, cuando un estudiante no se comporta adecuadamente, su participación es deficiente y perjudica al desenvolvimiento del grupo como equipo. La aparición de comportamientos individuales disfuncionales, se refleja de manera negativa en el rendimiento grupal e impiden alcanzar un aprendizaje adecuado [3]. El beneficio de la colaboración en el aprendizaje se logra a través del buen funcionamiento del equipo, esto es, cuando los alumnos interactúan entre sí, alentándose a preguntar, explicar y justificar sus ideas; cuando comparten información y conocimiento, muchas veces negociando para lograr un acuerdo, elaborando conocimiento y reflexionando sobre el mismo, y también cuando logran coordinar sus acciones [3, 4, 6]. Dado lo expuesto y desde una perspectiva psicosocial, es necesario que el alumno tenga desarrolladas habilidades colaborativas para poder comportarse de la manera adecuada [5].

En este artículo se analizó la manifestación de las habilidades de colaboración de los estudiantes, cuando utilizan herramientas sincrónicas y asincrónicas como integrantes de grupos de ACSC. Para ello se modelaron las interacciones usando oraciones de apertura, y se catalogó cada contribución de acuerdo con las habilidades de colaboración propuesta en [3]. Sin embargo, el propósito subyacente en esta investigación consiste en identificar el tipo de comunicación preferido por los estudiantes, a fin de que este conocimiento sea de utilidad para los docentes para propiciar procesos de enseñanza y de aprendizaje eficaces. Este trabajo forma parte de la investigación que está siendo desarrollada como Trabajo Final de Graduación de uno de los autores, en el marco del proyecto de investigación CICyT - UNSE Código 24/C097, titulado “Sistemas de información web basados en agentes para promover el Aprendizaje Colaborativo Soportado por Computadoras”.

El artículo se organiza de la siguiente manera. La sección 2 describe brevemente algunos trabajos considerados antecedentes de esta investigación. En la sección 3 se presenta la taxonomía de habilidades de colaboración que se utiliza en la investigación. En la sección 4 se menciona los tipos de comunicación, sus ventajas y

desventajas en relación al aprendizaje. En la sección 5, se presenta el concepto de interacción en entornos de ACSC, las dificultades en su análisis, y se introducen a las oraciones de apertura como técnica de modelación de interacciones. En la sección 6 se describe la experimentación realizada y se analizan los resultados obtenidos. Finalmente, en la sección 7 se enuncian algunas conclusiones.

2 Antecedentes

Existen algunos trabajos de investigación, donde se analizaron las interacciones de alumnos que aprenden colaborativamente, a través de herramientas de comunicación estructuradas o semiestructuradas, en entornos de ACSC. Baker y Lund son algunos de los pioneros en el intento de estructurar las interacciones de los estudiantes, y examinar las ventajas y desventajas del uso de las oraciones de apertura. Estos autores presentan en [7] un estudio comparativo de dos interfaces para la plataforma de ACSC, C-CHENE, que brinda soporte al aprendizaje colaborativo en tareas para resolver problemas de física. La primera interfaz consistió de un chat basado en texto, en donde los estudiantes se comunicaban sin restricciones mecanográficas con sus pares. Mientras que la segunda interfaz, se basa en una interfaz basada en oraciones de apertura, a la que los autores llaman "estructura flexible". Con esta estructura, sin embargo, los estudiantes están restringidos a seleccionar una oración predefinida para dialogar. Este estudio comprobó que una interfaz semiestructurada genera interacciones enfocadas en el problema planteado, evitando que los estudiantes desvíen la conversación hacia otros temas. Siguiendo esta línea de investigación, en [8] se analizan interacciones de estudiantes para comparar interfaces de comunicación en entornos de ACSC (semiestructuradas y con texto libre). Los autores concluyen que las manifestaciones relacionadas con tareas y estrategias se producen con más frecuencia a través de las interfaces semiestructuradas, mientras que las contribuciones de gestión se producen con más frecuencia a través de las interfaces libre. También se consideró interesante el trabajo de Balmaceda *et al.* [9]. En dicho trabajo se analizan las interacciones de grupo para detectar, de forma automática, roles en equipos de desarrolladores de software. Sin emnargo, cabe aclarar que en ninguno de los antecedentes mencionados se evalúan diferencias en el comportamiento colaborativo de los estudiantes, en base a la consideración del tipo de comunicación que utilicen, lo cual destaca la originalidad de este trabajo.

3 Habilidades de Colaboración

Según la clasificación o taxonomía creada por Soller [3], existen tres habilidades de colaboración que los estudiantes pueden manifestar: aprendizaje activo, conflicto creativo y conversación. Para cada una de dichas habilidades existen subhabilidades, y a su vez para cada una de estas, atributos que las describen. Esta clasificación fue estructurada desde la red de habilidades colaborativas ideada por McManus y Aiken [10], quienes a su vez se basaron en la investigación de Johnson *et al.* [11]. La Tabla

1 muestra las tres habilidades de colaboración con sus subhabilidades y atributos asociados.

Tabla 1. Taxonomía de habilidades del Aprendizaje Colaborativo [3]

Habilidad	Subhabilidad	Atributo	Oración de apertura
Conflicto Creativo	Mediación	Mediación Docente	*"Preguntémosle al profesor"*
	Argumentación	Conciliar	*"Ambos están correctos en eso"*
		Concertar	*"Yo estoy de acuerdo porque..."*
		Discrepar	*"Yo no estoy de acuerdo porque..."*
		Ofrecer alternativa	*"Alternativamente…"*
		Inferir	*"Entonces...", "Por lo tanto...."*
		Suponer	*"Si, ...entonces...."*
		Proponer excepciones	*"Pero podría ocurrir que"*
		Dudar	*"Yo no estoy seguro porque..."*
Aprendizaje Activo	Motivar	Animar	*"Muy Bien"*
		Reforzar	*"Está correcto"*
	Informar	Parafrasear	*"En otras palabras..."*
		Guiar	*"Yo pienso que deberían ..."*
		Sugerir	*"Yo pienso..."*
		Elaborar	*"Para elaborar.. " "Además..."*
		Explicar	*"Permítanme explicarlo ..."*
		Justificar	*"Para Justificar…"*
		Afirmar	*"Yo estoy seguro..."*
	Requerir	Información	*"¿Sabes tu...?"*
		Elaboración	*"¿Puedes decirme más?"*
		Clarificación	*"¿Puedes explicar cómo/por qué?"*
		Justificación	*"Por qué piensas eso"*
		Opinión	*"¿Piensas tu...?"*
		Ilustración	*"¿Por favor muéstrame?"*
Conversación	Reconocimiento	Apreciación	*"Gracias"*
		Aceptación/Confirmación	*"Bien" "Si"*
		Rechazo	*"No"*
	Mantenimiento	Requerir atención	*"Atiéndame..."*
		Sugerir acción	*"¿Podrías por favor....?"*
		Requerir confirmación	*"¿Está bien? "¿Es esto correcto?"*
		Atender	*"Yo te comprendo"*
		Disculparse	*"Discúlpame"*
	Tarea	Coordinar Procesos grupales	*"Bien, continuemos", "¿Están todos listos?"*
		Requerir cambio de enfoque	*"Permítanme mostrarles"*
		Resumir Información	*"Para resumir"*
		Finalizar participación	*"Adiós"*

4 Tipos de Comunicación

En [12] se dividen a los ambientes de ACSC en tres tipos o categorías de acuerdo con el tipo de comunicación que soportan: asíncronos, síncronos y multifunción, siendo los multifunción aquellos que soportan simultáneamente comunicación síncrona y

asíncrona. La comunicación asincrónica puede definirse como comunicación que se produce en cualquier momento y a intervalos irregulares, mientras que la comunicación síncrona es vista como cualquier comunicación que se produce en tiempo real [13]. Del conjunto de herramientas síncronas se puede mencionar al chat como una de las más utilizadas, y en el caso de las asíncronas al foro. A continuación se enuncian brevemente las ventajas y las desventajas de utilizar estos tipos de comunicación en entornos de ACSC.

Para un ambiente síncrono se considera ventajoso que la interacción entre los estudiantes sea inmediata y directa, permitiendo que ellos mismos regulen y monitoreen sus interacciones acorde con el contexto y la situación; la administración del proceso de aprendizaje sea dinámica y fluida, pues permite implementar nuevos tipos de tutoría dinámica y situada; posibilita el uso de nuevos recursos tecnológicos para interactuar, como jugar con identidades irreales (usando nicknames o personificando un avatar) o usar objetos virtuales. Debido a que las interacciones síncronas suelen ser más dinámicas y dialogadas, favorecen una construcción del conocimiento a través de negociación y consenso social, es decir, promueven la toma de decisiones colectivas. Sin embargo, pueden presentarse ciertas dificultades, por ejemplo, la coordinación de tiempos. Es así que suelen producirse inconvenientes cuando varios grupos deben conectarse al mismo tiempo y no se cuenta con la cantidad de profesores requerida para monitorear adecuadamente las actividades de todos. Además, como las comunicaciones síncronas suelen ser muy rápidas, los cambios de tema son dinámicos, y aunque todos los tópicos se encuentren vinculados, pierden actualidad rápidamente.

Para un ambiente asíncrono se observa positivamente que, por ejemplo, los alumnos interactúen cuando están preparados o tienen tiempo disponible para ello, permitiendo que reflexionen tanto tiempo como requieran, y que respondan a otros pares sólo cuando lo consideren apropiado. Sin embargo, podrían producirse demoras en la comunicación que impacten de manera negativa en los procesos de enseñanza y de aprendizaje. Por ejemplo, si un estudiante recibe respuesta a su comunicación luego de transcurrido un tiempo considerable puede suceder que haya olvidado el contexto dentro del que formuló su envío, o incluso puede pasar que las respuestas nunca lleguen [14]. Estas situaciones poco favorables dan origen a sentimientos de frustración, y también de soledad que no resultan beneficiosos para los procesos de monitoreo, tutoría y evaluación del aprendizaje [15].

5 Análisis de interacción en grupos de ACSC

Actualmente existen algunos cuestionamientos en relación al ACSC que deben ser investigados para mejorar la calidad de la enseñanza y del aprendizaje [16]. Algunas de estas cuestiones se relacionan estrechamente con el análisis de las interacciones en los campos del aprendizaje y el trabajo colaborativo soportado por computadoras (ACSC y TCSC, respectivamente), donde el esfuerzo investigativo se ha centrado en identificar y explorar los factores que afectan la eficacia y el éxito, del trabajo y del aprendizaje, en grupos que interaccionan en forma online [1].

El concepto general de interacción se enuncia como acción que se manifiestan entre dos o más individuos u objetos. Si bien toda interacción se inicia con una acción, es la reciprocidad de la misma la que establece si efectivamente se trata de una interacción [5]. Así, diversas teorías sostienen que son muy importantes factores como el nivel y la calidad de las interacciones en los procesos de aprendizaje. Esto se debe a que el conocimiento se construye activamente a través de la interacción [6].

En los últimos años se han realizado diferentes investigaciones sobre como analizar las interacciones de los grupos de alumnos que trabajan en forma colaborativa [16]. El fundamento que sostiene a dichas investigaciones, es que los alumnos que trabajan colaborativamente se comunican interactuando, entonces la colaboración implica interacción [17]. Por ende, un factor primordial para evaluar las habilidades de colaboración es contar con retroalimentación proveniente del análisis de las interacciones [18]. Obtener una retroalimentación que permita tomar decisiones correctivas en el comportamiento de los alumnos, implica necesariamente la observación y el análisis de las interacciones.

Una de las características de los entornos ACSC es que permiten la generación y almacenamiento de gran cantidad de datos acerca de los procesos de interacción y de ejecución de tareas por parte de los grupos de estudiantes [16, 19]. Sin embargo, una colaboración intensa, que incluya un número relativamente grande de interacciones hará que efectuar un seguimiento sea muy complicado, demandando demasiado tiempo y esfuerzo [16, 18, 20]. Aun cuando el tutor disponga de tiempo para realizar un análisis manual, es casi imposible que una única persona pueda detectar anomalías en las interacciones que puedan ocurrir en un grupo mediano de alumnos [21]. En [3] se menciona que el desarrollo de software para analizar la comunicación entre pares, es realmente una tarea importante, ya que las últimas tecnologías de comprensión del lenguaje natural combinadas en entornos de ACSC siguen siendo limitadas en su capacidad de comprender, e interpretar la comunicación del estudiante.

Una alternativa para facilitar el análisis de las interacciones consiste en utilizar una interfaz basada en oraciones de apertura. Una oración de apertura es una frase predefinida que se utiliza para comenzar una contribución en un diálogo. Con este tipo de representación el usuario está obligado a elegir, desde una lista de frases, aquella que mejor indique la intención de su colaboración. Generalmente se implementan en interfaces gráficas en forma de menús con botones, y se brinda al estudiante la opción de completar su mensaje en un área de texto libre [4, 5]. Aunque se encuentren opiniones antagónicas sobre los beneficios de utilizar las oraciones de apertura para modelar las interacciones, existen investigaciones que demuestran que las mismas facilitan la comunicación, y simplifican la identificación y el análisis de secuencias de interacciones conversacionales en entornos de ACSC [4, 5].

6 Experimentación y análisis de resultados

Con fines de experimentación se desarrollaron dos herramientas de comunicación, un chat y un foro. Para facilitar el estudio de las contribuciones de los alumnos que utilizarían estas herramientas, se decidió crear para ambos desarrollos interfaces semiestructuradas basadas en oraciones de apertura. El conjunto de oraciones de

apertura implementado en cada interfaz fue el mismo y respondió a la clasificación de habilidades colaborativas propuesta por Soller [3]. Esto quiere decir que, cuando un alumno desee realizar una contribución, debió escoger una oración de apertura en la interfaz semiestructurada para luego completar su contribución con texto libre. Considerando la correspondencia existente entre los atributos de colaboración y las oraciones de apertura disponibles, se calcularon indicadores que permitieron determinar cómo manifiestan los estudiantes sus habilidades de colaboración cuando trabajan en forma síncrona, y también cuando lo hacen de forma asíncrona.

La experimentación con las herramientas de comunicación y colaboración desarrolladas, se realizó contando con 56 (cincuenta y seis) estudiantes de las carreras Licenciatura en Sistemas de Información y Programador Universitario en Informática, ambas pertenecientes a la UNSE. Con treinta y dos de estos estudiantes, once mujeres y veintiún varones, se organizaron 11 (once) grupos que trabajaron utilizando la herramienta asíncrona, es decir, el foro basado en oraciones de apertura. Los estudiantes fueron asignados aleatoriamente como integrantes de los grupos, resultando un grupo de dos integrantes y diez grupos de tres. Los veinticuatro estudiantes restantes, quince mujeres y nueve varones, se distribuyeron en 7 (siete) grupos que trabajaron utilizando la herramienta síncrona, es decir, el chat basado en oraciones de apertura. En este caso los estudiantes también fueron asignados aleatoriamente como integrantes de los grupos, resultando cuatro grupos de cuatro integrantes, un grupo de dos y dos grupos de tres.

Las sesiones síncronas utilizando el chat basado en oraciones de apertura se llevaron a cabo en el Laboratorio Alfa del Departamento de Informática de la FCEyT de la UNSE, para así poder constatar que los estudiantes efectivamente colaboraban en tiempo real. La duración de las sesiones fue aproximadamente 150 minutos. Los estudiantes fueron distribuidos en el laboratorio de manera tal que no fuera posible ningún contacto presencial entre integrantes de un mismo grupo, asegurando que el diálogo sólo pudiera efectuarse mediante el uso de la herramienta.

Las experiencias asíncronas utilizando el foro basado en oraciones de apertura, se desarrollaron en el lapso de tiempo máximo de una semana desde las locaciones y en los momentos en que los estudiantes consideraron adecuados.

Considerando los resultados obtenidos, el primer aspecto a resaltar es que el total de interacciones registradas ascendió a 1.256, siendo 598 contribuciones realizadas en el foro basado en oraciones de apertura y 658 en el chat. La Tabla 3 muestra la cantidad de contribuciones para cada tipo de habilidad y los porcentajes correspondientes, discriminando los resultados según forma de trabajo, es decir, síncrona y asíncrona. En la Tabla mencionada puede visualizarse que en ambos tipos de comunicación la habilidad de “Conversación” es la más utilizada. Sin embargo, no hubo coincidencia en la segunda habilidad más utilizada, resultando el “Aprendizaje activo” en las sesiones síncronas y “Conflicto creativo” en las asíncronas, ambas con porcentajes de uso casi iguales. Prácticamente puede expresarse lo mismo respecto a la habilidad menos utilizada en ambas sesiones, ya que en las síncronas fue la habilidad de “Conflicto creativo” y en las asíncronas “Aprendizaje activo”.

Considerando los resultados detallados en Tabla 4, respecto a la manifestación de subhabilidades y atributos de la habilidad “Conflicto creativo”, se observa supremacía de la subhabilidad “Argumentación”. La mayor frecuencia de aparición en esta subhabilidad, tanto en forma síncrona como asíncrona, responde al atributo “Inferir”,

siguiendo en segundo lugar “Concertar” y en tercero “Discrepar”. Los demás atributos de la subhabilidad presentan bajos porcentajes de aparición. También, se observan escasas muestras de la subhabilidad “Mediación”.

Tabla 3. Cantidad de interacciones discriminando por tipo de sesión.

Habilidad de colaboración	*Contribuciones en sesiones síncronas*	*Porcentajes en sesiones síncronas*	*Contribuciones en sesiones asíncronas*	*Porcentajes en sesiones asíncronas*
Aprendizaje Activo	226	34.34	156	26.09
Conflicto Creativo	183	27.81	212	35.45
Conversación	249	37.85	230	38.46
Total	**658**	**100**	**598**	**100**

Tomando la habilidad “Aprendizaje activo”, la mayor frecuencia de aparición está asociada a la subhabilidad “Informar” en ambos tipos de sesiones. El atributo de colaboración “Sugerir” coincidentemente también es el más utilizado en ambos casos. Con cantidades inferiores aparece el resto de atributos de la mencionada subhabilidad, observándose distribuciones de aparición similares en ambos tipos de sesiones.

Para la habilidad de colaboración “Conversación”, en ambos tipos de sesiones se muestra en supremacía la subhabilidad “Reconocimiento”, para su atributo “Aceptación/Confirmación”. El resto de subhabilidades y atributos presentan porcentajes inferiores distribuidos de manera similar en ambos tipos de sesiones.

7 Conclusiones

En el presente artículo se presentan los resultados obtenidos a través de un estudio observacional sobre interacciones de grupos de estudiantes universitarios de informática. Mediante experimentación se recopilaron y analizaron 1.256 contribuciones, 598 pertenecientes a comunicaciones asíncronas y 658 a síncronas. Esta diferencia muestra un mayor número de manifestaciones de habilidades de colaboración cuando trabajaban con la herramienta síncrona que cuando lo hicieron con la asíncrona. Dado que en la experimentación la cantidad de grupos operativos asíncronos fue considerablemente mayor que los grupos síncronos, podría considerarse que esta diferencia a favor de las contribuciones síncronas estaría mostrando un indicio respecto a la preferencia de trabajo de los estudiantes lo cual podría ser considerado por los docentes al momento de planificar sus actividades y definir el tipo de herramienta a utilizar. Por otro lado, los resultados también indican que los estudiantes tienen una mayor tendencia a manifestar habilidades de “Aprendizaje activo” cuando colaboran de forma asíncrona, mientras que la habilidad de “Conflicto creativo” tiene mayor manifestación cuando lo hacen de forma asíncrona. Este conocimiento podría ser útil para los docentes al momento de determinar el tipo de actividad a desarrollar por los estudiantes. Curiosamente, la habilidad de “Conversación”, la cual tiene mayor ocurrencia en ambos tipos de comunicación, tiene también un nivel de manifestación similar, con lo cual se podría

suponer que es la habilidad menos afectada por el tipo de comunicación empleado por los estudiantes al momento de colaborar.

En base a éste análisis inicial se puede concluir que existen ciertas diferencias en la manifestación de las habilidades de colaboración dependiendo del tipo de comunicación que utilicen los estudiantes, pero las mismas no son substanciales.

Tabla 4. Resultados considerando subhabilidades y atributos de colaboración

Atributo de Colaboración	**Cantidad de muestras en sesiones asíncronas**	**%**	**Subhabilidad/ N° de manifestaciones en asíncrono**	**Cantidad de muestras en sesiones asíncronas**	**%**	**Subhabilidad/N° de manifestaciones en síncrono**
Mediación Docente	3	0.5	Mediación/ 3	3	0.45	Mediación/ 2
Conciliar	1	0.17	Argumentación/ 209	5	0.76	Argumentación/ 181
Concertar	20	3.34		34	5.16	
Discrepar	9	1.5		22	3.34	
Ofrecer alternativa	1	0.17		0	0	
Inferir	157	26.25		104	15.8	
Proponer excepción	4	0.66		5	0.76	
Suponer	5	0.84		3	0.45	
Dudar	12	2		8	1.21	
Animar	3	0.5	Motivar/ 14	7	1.06	Motivar/ 19
Reforzar	11	1.84		12	1.82	
Parafrasear	18	3.01	Informar/ 120	35	5.31	Informar/ 173
Guiar	15	2.59		25	3.8	
Sugerir	70	11.7		80	12.15	
Elaborar	3	0.5		2	0.3	
Explicar	8	1.34		19	2.88	
Justificar	1	0.17		2	0.3	
Afirmar	5	0.84		9	1.36	
Información	6	1	Requerir/ 22	12	1.82	Requerir/ 34
Elaboración	4	0.66		11	1.67	
Clarificación	8	1.34		3	0.45	
Justificación	0	0		2	0.3	
Opinión	3	0.5		3	0.45	
Ilustración	1	0.17		3	0.45	
Apreciación	6	1	Reconocimiento/ 97	14	2.12	Reconocimiento/ 102
Aceptación/Confirmación	86	14.38		82	12.46	
Rechazo	5	0.84		6	0.91	
Requerir atención	1	0.17	Mantenimiento/ 62	4	0.6	Mantenimiento/ 74
Sugerir acción	8	1.34		7	1.06	
Requerir confirmación	21	3.51		22	3.34	
Atender	6	1		6	0.91	
Disculparse	26	4.35		35	5.31	
Coordinar Procesos grupales	35	5.85	Tarea/ 71	35	5.31	Tarea/ 73
Requerir cambio de enfoque	1	0.17		1	0.15	
Resumir Información	16	2.67		22	3.34	
Finalizar participación	19	2.59		15	2.28	
Total	**598**	**100**	**598**	**658**	**100**	**658**

Referencias

1. Dillenbourg, P.: What do you mean by collaborative leraning?. In P. Dillenbourg (Ed) Collaborative-learning: Cognitive and Computational Approaches, 1–19. Oxford, Elsevier (1999)
2. Wang, Qiyun: Design and evaluation of a collaborative learning environment. Learning Sciences and Technologies Academic Group, National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore 637616, Singapore (2009)
3. Soller, A. L.: Supporting social interaction in an intelligent collaborative learning system. International Journal of Artificial Intelligence in Education, 12, 40–62 (2001)
4. Lazonder, A.W., Wilhelm, P., Ootes, S.A.W.: Using Sentence Openers to Foster Student Interaction in Computer-Mediated Learning Environments. Computers & Education, 41(3), 291–308 (2003)
5. Costaguta, Rosanna Nieves: Entrenamiento de Habilidades Colaborativas. Facultad de Ciencias Exactas, Departamento de Computación y Sistemas, Universidad Nacional del Centro de la Pcia. De Bs. As., Mayo 2008. (2008)
6. Orvis, Kara L., Lassiter, Andrea L. R.: Computer-Supported Collaborative Learning: The Role of the Instructor. En: Ferris Sharmila Pixy-y Godar Susan H (eds). Teaching and Learning with Virtual Teams. USA, Information Science Publishing, pp. 158–179 (2006)
7. Baker, M., Lund, K.: Promoting reflective interactions in a computer-supported collaborative learning environment. Journal of Computer-Assisted Learning, 12, 17–193 (1997)
8. Jermann, P., Schneider, D.K.: Semi-structured interface in collaborative problem-solving. Swiss workshop on collaborative and distributed systems. Lausanne, May 1st. (1997)
9. Balmaceda, J., García, P., Schiaffino, S.: Detección Automática de Roles de Equipo. Proceedings of the X Simposio Argentino de Inteligencia Artificial (39 JAIIO), 235–238 (2010)
10. McManus, M., Aiken, R.: Teaching collaborative skills with a group leader computer tutor. Education and Information Technologies, 1, 75–9 (1996)
11. Johnson, D., Johnson, R., Holubec, E.: Circles of learning: Cooperation in the classroom (3ra Ed.). Edina, MN: Interaction Book Company (1990)
12. Qu, C., Nejdl, W.: Constructing a Web-Based Asybchronous and Synchronous Collaboration Environment Using WebDAV and Lotus SameTime. In Proceedings of the ACM SIGUCCS, Portland, Oregon, USA. (2001)
13. Kligyte, G., Leinonen, T.: Study of functionality and interfaces of existing CSCL/CSCW systems, Unpublished notes. http://www.euro-cscl.org/site/tcole/ublic_eliverables_html (2001)
14. Fahraeus, E., Chamberlain, B., Bridgeman, N., Fuller, U., Rugelj, J.: Teaching with Electronic Collaborative Learning Groups. In ITiCSE'99 Working Group Reports, 31 (4), 121–128 (1999)
15. King, F.: A virtual student Not an ordinary Joe. Internet and Higher Education, 5, 157–166 (2002)
16. Daradoumis, T., Martinez, A., Xhafa, F.: A Layered Framework for Evaluating Online Collaborative Learning Interactions. International Journal of Human-Computer Studies. 64(7), 622–635 (2006)
17. Costaguta, Rosanna Nieves: Habilidades de Colaboración Manifestadas por los Estudiantes de Ciencias de la Computación. Revista Nuevas Propuestas. Universidad Catolica de Santiago del Estero, Argentina. Vol. 43-44 (2008)
18. Chen, Weiqin, Wasson, Barbara: An Instructional Assistant Agent for Distributed Collaborative Learning. En S.A. Cerri, G. Gouardères, y F. Paraguaçu (eds) Intelligent Tutoring Systems, Lecture Notes in Computer Science: Proceeding of ITS 2002, LNCS 2363, pp. 609–618. Alemania, Springer-Verlag (2002)
19. Barros, B., Verdejo, M. F.: Analysing student interaction processes in order to improve collaboration: the degree approach, International Journal of Artificial Intelligence in Education, 11, 221–241 (2000)
20. Rosé, C., Wang, Yi-Chia, Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning, 3 (3), 237–271 (2008)
21. Daradoumis, T., Martínez, A., Xhafa, F.: An Integrated Approach for Analysing and Assessing the Performance of Virtual Learning Groups. In: de Vreede, G.-J., Guerrero, L.A., Marín Raventós, G. (eds.) CRIWG 2004. LNCS, vol. 3198, pp. 289–304. Springer, Heidelberg (2004)

X WORKSHOP INGENIERÍA DE SOFTWARE

- WIS -

CACIC2013

X WORKSHOP INGENIERÍA DE SOFTWARE

- WIS -

ID	Trabajo	Autores
5595	**Reingeniería de una Línea de Productos de Software: Un Caso de Estudio en el Subdominio de Ecología Marina**	Natalia Huenchuman (UNCOMA), Agustina Buccella (UNCOMA), Alejandra Cechich (UNCOMA), Matias Pol'la (UNCOMA), Maria del Socorro Doldan (UNCOMA), Enrique Morsan (UNCOMA), Maximiliano Arias (UNCOMA)
5867	**Generación Automática del Modelo de Diseño desde el Modelo de Análisis a través de Reglas QVT**	Ariel Arsaute (UNRC), Fabio Zorzan (UNRC), Marcela Daniele (UNRC), Paola Martellotto (UNRC)
5617	**Modeling Complex Mobile Web Applications from UI Components? Adding Different Roles and complex Database Design**	Pablo Vera (UNLaM), Claudia Pons (UNLP), Carina González (ULL), Daniel Giulianelli (UNLaM), Rocío Rodríguez(UNLaM)
5671	**Un Análisis Experimental de Tipo de Aplicaciones para Dispositivos Móviles**	Lisandro Delia (UNLP), Nicolás Galdámez (UNLP), Pablo Thomas (UNLP), Patricia Pesado (UNLP)
5668	**Evolución Semántica de Glosarios en los Procesos de Requisitos**	Gladys Kaplan (UNLaM), Jorge Doorn (UNCPBA), Nora Gigante (UNLaM)
5692	**Generación de un Algoritmo de Ranking para Documentos Científicos del Área de las Ciencias de la Computación**	H. Kuna (UNaM), M. Rey (UNaM), J. Cortes (UNaM), E. Martini (UNaM), L. Solonezen (UNaM), R. Sueldo (UNaM)
5723	**Gestión de Contenido Organizacional (ECM)**	Alejandra López (UNPSJB), Julio Moreyra (UNPSJB)
5700	**Marcos Metodológicos dentro de la Informática**	Pablo J. Iuliano (UNLP), Luis Marrone (UNLP), Elvio Fernandez (UNLP)

X WORKSHOP INGENIERÍA DE SOFTWARE

- WIS -

ID	Trabajo	Autores
5703	**Agentes Inteligentes para propiciar la Accesibilidad Web**	Gabriela Miranda (UNPA), Adriana Martin (GIISCO), Rafaela Mazalú (UNCOMA), Gabriela Gaetan (UNPA), Viviana E. Saldaño (UNPA)
5755	**Evaluación de Accesibilidad del Contenido Web Utilizando Agentes**	Rafaela Mazalú (UNCOMA), Alejandra Cechich (UNCOMA), Adriana Martín (UNPA)
5701	**QUCO2: Development of a tool for measuring the quality of Web applications**	Nicolás Tortosa (UTN-FRRe), Noelia pinto (UTN-FRRe), Cesar Acuña (UTN-FRRe), Liliana Raquel Cuenca Pletsch (UTN-FRRe), Marcelo Estayno (UNLZ)
5741	**Análisis de la información presente en foros de discusión técnicos**	Nadina Martinez, Gabriela Aranda (UNCOMA), Mauro Sagripanti (UNCOMA), Pamela Faraci (UNCOMA), Alejandra Cechich (UNCOMA)
5858	**Knowledge Management in Distributed Software Development: A Systematic Review**	Fernanda Tamy Ishii , Gislaine Camila L. Leal, Edwin V. Cardoza Galdamez, Elisa Hatsue M. Huzita, Renato Balancieri, Tânia Fátima C. Tait
5731	**Propuesta de una Metodología para el Análisis de Adopción de Cloud Computing en PyMEs**	Luciano Bernal (UTN-FRBA), Cinthia Vegega (UTN), Pablo Pytel (UNLA), Maria Florencia Pollo Cattaneo (UTN-FRBA)
5844	**Modelo para aplicaciones sensibles al contexto (MASCO): Un caso de estudio para validación**	Evelina Carola Velazquez (UNJu), Ariel Nelson Guzmán Palomino (UNJu), María del Pilar Galvez Díaz (UNJu), Nélida Raquel Caceres (UNJu)
5845	**Um mecanismo de captura de informações contextuais em um Ambiente de Desenvolvimento Distribuído de Software**	Yoji Massago, Renato Balancieri, Raqueline Ritter de Moura Penteado, Elisa Hatsue Moriya Huzita, Rafael Leonardo Vivian

X WORKSHOP INGENIERÍA DE SOFTWARE

- WIS -

ID	Trabajo	Autores
5690	**Q-Scrum: una fusión de Scrum y el estándar ISO/ IEC 29110**	Ariel Pasini (UNLP), Silvia Esponda (UNLP), Marcos Boracchia (UNLP), Patricia Pesado (UNLP)
5719	**Inserción del mantenimiento en los procesos ágiles**	Karla Mendes Calo (UNS), Karina M. Cenci (UNS), Pablo Rubén Fillottrani (UNS)
5813	**Trazabilidad de Procesos Ágiles: un Modelo para la Trazabilidad de Procesos Scrum**	Roberto Nazareno (UNLAR), Silvio Gonnet (UTN), Horacio Leone (UTN)
5759	**Evaluación de variantes en modelo destinado a anticipar la conveniencia de trazar proyectos de software**	Juan Giró (UTN), Juan Carlos Vazquez (UTN-FRC), Brenda E. Meloni (UTN-FRC), Leticia Constable (UTN-FRC)
5760	**Análisis de rendimiento del algoritmo SGP4/ SDP4 para predicción de posición orbital de satélites artificiales utilizando contadores de hardware**	Federico Diaz (UNLaM), Fernando G. Tinetti (UNLP), Nicanor Casas (UNLaM), Sergio Martín (UNLaM), Graciela De Luca (UNLaM), Daniel Giulianelli (UNLaM)

Reingeniería de una Línea de Productos de Software: Un Caso de Estudio en el Subdominio de Ecología Marina

Natalia Huenchuman[1] *, Agustina Buccella[12], Alejandra Cechich[1], Matias Pol'la[12], Maria del Socorro Doldan[3], Enrique Morsan[3], and Maximiliano Arias[1]

[1] GIISCO Research Group
Departamento de Ingeniería de Sistemas - Facultad de Informática
Universidad Nacional del Comahue
Neuquen, Argentina
natalia.huenchuman@gmail.com,{agustina.buccella,alejandra.cechich,matias.polla}@fi.uncoma.edu.ar,
ariasmaxi89@gmail.com
[2] Consejo Nacional de Investigaciones Científicas y Técnicas - CONICET
[3] Instituto de Biología Marina y Pesquera "Almirante Storni"
Universidad Nacional del Comahue - Ministerio de Producción de Rio Negro
San Antonio Oeste, Argentina
{msdoldan,qmorsan}@gmail.com

Abstract. La ingeniería de líneas de productos de software y la ingeniería de software basada en componentes persiguen objetivos similares: minimizar los costos y el esfuerzo en el desarrollo de nuevos sistemas utilizando al reuso como la herramienta principal. Aunque poseen bases diferentes en cuanto a la manera de llevar a cabo un desarrollo de software, ambas ingenierías pueden ser combinadas para maximizar el reuso efectivo e implementar sistemas de alta calidad en un menor tiempo. En este trabajo se presenta la reestructuración de una línea de productos creada para el subdominio de Ecología Marina para orientarla estrictamente a componentes reusables (utilizando herramientas de código abierto).

Keywords: Reingeniería de Software, Sistemas de Información Geográficos, Líneas de Productos de Software, Ecología Marina, Herramientas de Código Abierto

1 Introduccón

La Ingeniería de Software Basada en Componentes (ISBC) [7, 17] provee metodologías, técnicas y herramientas basadas en el uso y ensamblaje de piezas pre-fabricadas (desarrolladas en momentos diferentes, por distintas personas y posiblemente con distintos objetivos de uso) que puedan formar parte de nuevos sistemas a desarrollar. El objetivo final de esta ingeniería es minimizar los costos y tiempo de desarrollo de los sistemas, incluso mejorando la calidad de los mismos. A pesar de que la tecnología de componentes de software ha comenzado hace ya un largo tiempo atrás, aproximadamente desde la década del 60 [12], todavía hay aspectos que siguen siendo analizados para obtener beneficios tangibles al desarrollar nuevos sistemas. A su vez, han surgido nuevos paradigmas que poseen objetivos similares a los de la ISBC, pero que se enfocan en aspectos diferentes a la hora de construir sistemas. Uno de estos paradigmas, fuertemente analizado en la actualidad, es la Ingeniería de Líneas de Productos de Software (ILPS) [3, 5, 15, 18] la cual provee mecanismos para la definición de activos comunes junto con una variabilidad controlada dentro de un dominio en particular. Las principales diferencias entre ambos enfoques, líneas de productos y componentes, radican en dos aspectos principales. En primer lugar, en las líneas de productos de software se reusan los activos que fueron diseñados explícitamente para su reutilización, es decir, se crean con un objetivo de predictibilidad contra el oportunismo del desarrollo de componentes. En segundo lugar, las líneas de productos son tratadas como un todo, no como productos múltiples que se ven y se mantienen por separado como en el desarrollo de componentes[4]. Sin embargo, a pesar de que en el enfoque de líneas de productos no se obliga explícitamente al desarrollo de los activos utilizando componentes de software, el uso combinado de ambos enfoques contribuye a reducir el

* Este trabajo esta parcialmente soportado por el proyecto UNCOMA F001 "Reuso Orientado a Dominios" como parte del programa "Desarrollo de Software Basado en Reuso".

[4] A Framework for Software Product Line Practice, Version 5.0. http://www.sei.cmu.edu/productlines/frame_report/pl_is_not.htm

acoplamiento e incrementar la cohesión, mejorando la modularidad y la evolución de los sistemas construídos [1, 2].

En trabajos anteriores [13, 14] se realizó una línea de productos de software (LPS) aplicada al subdominio de Ecología Marina que sirvió como una plataforma de servicios comunes, aplicables a todos los productos de la línea, y servicios variables, aplicados solo a algunos productos. Dicha línea fue creada a partir de la metodología de desarrollo de LPS orientada a subdominios [14] la cual combinaba ventajas de otras metodologías muy referenciadas tanto en el ámbito académico como en la industria [3, 6, 10, 15]. A pesar de que la LPS sirvió en un primer momento para crear un producto dentro del dominio (creado para el Instituto de Biología Marina y Pesquera "Almirante Storni"[5]), surgieron varios inconvenientes a la hora de realizar modificaciones o crear nuevos productos para otras organizaciones; el código se encontraba altamente acoplado y no permitía un reuso sencillo y real de la plataforma y de la instanciación de la variabilidad. Para solucionar este inconveniente se buscó entonces rediseñar la LPS aplicando técnicas que permitan un desarrollo de componentes de software para lograr así un reuso efectivo dentro del dominio geográfico. Esto generó que no sólo fuera necesario un cambio en la tecnología en que estaba implementada la LPS, sino también cambios en el enfoque de desarrollo; lo que llevó a un proceso de reestructuración de la LPS. Dicho proceso incluyó parte de reestructuración de código, debido a que los componentes fueron reescritos en otro lenguaje; e ingeniería directa, debido a que se modificó el diseño de la línea agregando nuevos requerimientos y modificando los previos.

En la literatura existen varios trabajos que proponen nuevas metodologías o presentan casos de estudio que reestructuran lineas de productos de software hacia un enfoque de componentes [5, 9, 11, 16, 19]. Por ejemplo, en [9] se presenta la aplicación de una metodología para reestructurar un sistema de soporte de inventarios de negocios. A su vez, en [16], se presenta una aplicación del método de Análisis de Opciones para la Reingeniería (OAR) centrado en la minería de componentes existentes de una línea de productos o de una nueva arquitectura de software. Por último, en [11] se realiza una comparación de las propuestas más referenciadas en la literatura enfocándose en los nuevos desafíos respecto al área de reestructuración o reingeniería de líneas de productos. En este trabajo utilizamos una metodología que combina algunas de estas propuestas con nuestra metodología enfocada en subdominios de aplicación [13, 14].

Este artículo está organizado de la siguiente manera. A continuación se describe la metodología utilizada para la reestructuración de una LPS en el dominio geográfico junto con las actividades involucradas para el análisis, diseño y desarrollo de componentes reusables. La metodología se presenta mediante la ilustración de un caso de estudio real en donde se pueden ver los pasos realizados para la reconstrucción de la LPS creada para el subdominio de ecología marina, y de los beneficios obtenidos. Finalmente en la última sección se presentan las conclusiones y trabajo futuro.

2 Reestructuración de la Línea de Productos de Software

Para llevar adelante la reestructuración de la LPS, se combinó el método de *Análisis de Opciones para la Reingeniería* (OAR), presentado en [5] junto con metodología propuesta para el dominio de ecología marina, presentada en [4, 14]. El método OAR fue seleccionado debido a las experiencias de su aplicación [5, 16] y la viabilidad de aplicarla a la LPS heredada.

En la Figura 1 se muestran las siete actividades de la metodología correspondientes a la fase de *Ingeniería de Dominio* del desarrollo de la LPS en donde se puede ver la influencia de los estándares para la realización de algunas de ellas. A continuación se describen las tareas involucradas en estas siete actividades junto con los trabajos que se han llevado a cabo para aplicarlas al caso de estudio dentro del subdominio de ecología marina:

- *Análisis de Documentación e Información:* esta actividad incluye principalmente adquirir el conocimiento del dominio analizado, los ámbitos donde tiene alcance y el estudio de la documentación brindada (artefactos de software plasmados como modelos de variabilidad, arquitectura de referencia, modelo conceptual, etc.). También, se deben considerar los beneficios que

[5] http://ibmpas.org/

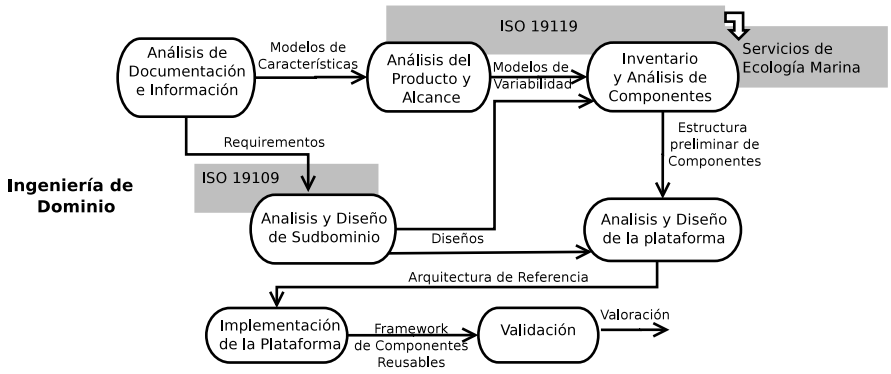


Fig. 1. Metodología para reestructurar la LPS

aporta a la organización el producto implementado y los requerimientos que no fueron relevados anteriormente.

Para realizar estas tareas, se analizó nuevamente el dominio de ecología marina mediante la información provista por referentes expertos en dichos dominios. Con esta información se pudieron analizar los requerimientos iniciales contra los nuevos resultando en la adición de dos nuevas características, la ejecución de determinadas consultas espaciales y la generación de datos estadísticos. Dichos requerimientos no fueron contemplados en el diseño anterior de la LPS, por lo que debieron ser agregados a la nueva versión. La descripción detallada del diseño original de la LPS puede verse en [13, 14].

– *Análisis del Producto y Alcance:* esta actividad se realiza junto al análisis obtenido de la etapa anterior que sirve como guía aportando la funcionalidad y el comportamiento propio del negocio, las decisiones de diseño e implementación y los nuevos requerimientos que se buscan cumplir. Aquí se deben estudiar las herramientas de software que fueron utilizadas, analizar las nuevas características, e investigar y seleccionar las nuevas herramientas de software a utilizar que cumplan con los requisitos solicitados y se adecúen a las reglas definidas en los estándares propuestos por el Consorcio Geo-Espacial (OGC)[6] y el Comité Técnico ISO/TC 211[7]; en particular de la normas de Arquitectura de Servicios (OpenGIS Service Architecture)[8] y de la ISO 19119[9].

  En nuestro trabajo, primero se analizó la LPS heredada debido a los problemas surgidos al momento de reusarla; su diseño altamente acoplado generaba que el reuso de la funcionalidad sea muy dificil de realizar tomando mucho tiempo y esfuerzo crear nuevos productos de la línea. En general, estos inconvenientes se generaban básicamente debido a la tecnología empleada para el desarrollo de la plataforma. La arquitectura de la LPS heredada fue desarrollada a nivel de módulos utilizando librerías de código abierto para la gestión de datos geográficos. Las mismas fueron: PostGIS[10], para la creación de la base de datos espacial; Geoserver[11], como servidor de mapas para publicar datos a partir de las principales fuentes de datos espaciales que utilizan estándares abiertos; y OpenLayers[12], como cliente ligero web, elegido debido a la facilidad de acceso a su configuración y adaptación del código fuente. A pesar de que muchas de estas herramientas son muy utilizadas en aplicaciones actuales, no permiten desarrollos independientes que incrementen la modificabilidad y evolución. Por ejemplo, podemos afirmar que la herramienta OpenLayers lidera el ranking de uso en aplicaciones geográficas de código abierto en la actualidad. Sin embargo, a pesar de ser fácil de usar, posee muchos inconvenientes derivados de su tecnología similar a JavaScript. Al ejecutarse del lado del cliente genera que

[6] `http://www.opengeospatial.org/`
[7] `http://www.isotc211.org/`
[8] The OpenGIS Abstract Specification: Service Architecture, 2002.
[9] Geographic information. Services International Standard 19119, ISO/IEC,2005.
[10] `http://postgis.refractions.net/`
[11] `http://geoserver.org/display/GEOS/Welcome`
[12] `http://openlayers.org/`

haya que considerar el código dependiendo las particularidades de cada navegador, y además su bajo nivel de tipado también genera varias inconsistencias. Además, el inconveniente más importante es que toda la lógica del negocio debe ser incluída en la interface de cada módulo generando un alto acoplamiento en el código que imposibilita el reuso efectivo de los servicios de la plataforma. Este análisis de las herramientas utilizadas en la LPS heredada y de las nuevas herramientas a aplicar estuvo guiado por la capacidad de las mismas de permitir la implementación de componentes que soporten el desarrollo de la arquitectura definida según los estándares geográficos. Dicha arquitectura, de al menos 3 niveles, promueve la separación de la interface, en una capa de *interface de usuario* que es responsable de la interacción con el usuario; de la lógica del negocio, en una capa de *procesamiento geográfico* responsable de coordinar e implementar la funcionalidad requerida por el usuario; y de la administración y almacenamiento de los datos geográficos, en una capa de *modelado geográfico* responsable del almacenamiento de los datos físicos y su manipulación. La principal ventaja de esta arquitectura es la separación de la funcionalidad en tres capas independientes que interactúan a través de sus interfaces bien definidas. Considerando entonces esta arquitectura, y luego de un estudio exhaustivo de las posibles herramientas de código abierto disponibles en la Web [8], se eligieron principalmente cinco de ellas encargadas de implementar los requerimientos de cada capa de la arquitectura. La Figura 2 muestra dichas herramientas dentro de la capa que deben implementar.

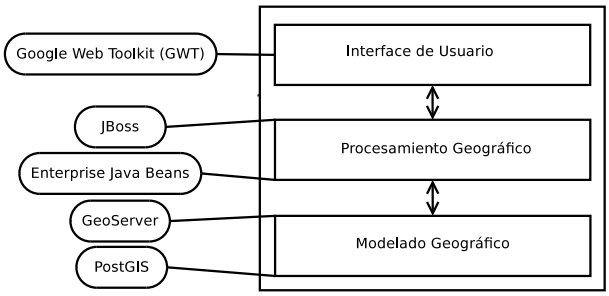


Fig. 2. Herramientas de código abierto seleccionadas para gestionar cada capa

Como vemos, para la capa de *modelado geográfico* se decidió emplear el mismo sistema gestor de base de datos espaciales *PostGIS* que en la LPS heredada [14], ya que posee una serie de ventajas que lo posicionan primero entre las opciones de este tipo de software. Entre ellas se destacan que es software libre, tiene licencia GNU General Public License (GPL), es compatible con los estándares de OGC, soporta tipos espaciales, índices espaciales, etc. Dentro de la misma capa, con respecto al servidor geográfico también se decidió mantener el mismo, *Geoserver*, principalmente, por la mantenibilidad, debido a que es muy simple manejar la configuración a través de una interface web amigable. Una vez seleccionadas las herramientas iniciales, se comenzó la búsqueda de las herramientas para el desarrollo de componentes, los cuales son implementados como parte de la capa de *procesamiento geográfico*. Entre las variadas herramientas existentes, se seleccionó la tecnología Enterprise Java Beans (EJB)[13] debido a su amplia utilización por simplificar el proceso de creación de componentes permitiendo al programador abstraerse de los problemas generales de una aplicación (concurrencia, transacciones, persistencia, seguridad, etc.) para centrarse en el desarrollo de la lógica de negocio en sí. Luego, para el despliegue de los EJBs se eligió al servidor *JBoss*[14]. El mayor desafío fue luego la selección de las herramientas para crear la interface web de usuario, es decir, los clientes web de servicios geográficos (correspondientes a la capa de *interface de usuario*), ya que existen en la Web decenas de estas herramientas cada una con diferentes características. Por ejemplo, algunas de ellas usan únicamente tecnología del lado del cliente mientras que la amplia mayoría depende de funcionalidades del lado del servidor. Así permiten la ejecución de tareas avanzadas como seguridad, administración de usuarios y grupos, análisis espacial y personalización de controles

[13] Oracle, http://www.oracle.com/technetwork/java/javaee/ejb/index.html
[14] `http://www.jboss.org`

y funcionalidades de interfaces gráficas de usuario, entre otras. Afortunadamente, el OGC ha promovido el uso de estándares para servicios web de mapas que han ayudado a establecer un marco común de trabajo para acceder a información geográfica en internet (Web Map Service, Web Feature Service, Web Coverage Service), presentarla por medio de estilos (Style Layer Descriptor), filtrarla (Filter encoding), almacenarla, transportarla (Geography Markup Language y Keyhole Markup Language) y procesarla (Web Processing Service). A su vez, muchos de los clientes web existentes tienen dependencias entre ellos, algunos han desaparecido y otros se toman como base de nuevos desarrollos. En la Figura 3 se muestra dicha dependencia entre los clientes web para GIS extraída de OSGeo[15]. En este trabajo, se utilizó ese conjunto de herramientas como conjunto inicial y se las clasificó de acuerdo a tres aspectos principales: cuáles de ellas son proyectos oficialmente abandonados (marcados con un triángulo rojo), cuáles no poseen una versión reciente (identificados con triángulo amarillo), es decir que llevan más de un año sin una nueva versión, y cuáles pueden ser consideradas como vigentes hoy en día (indicadas por un tilde verde).

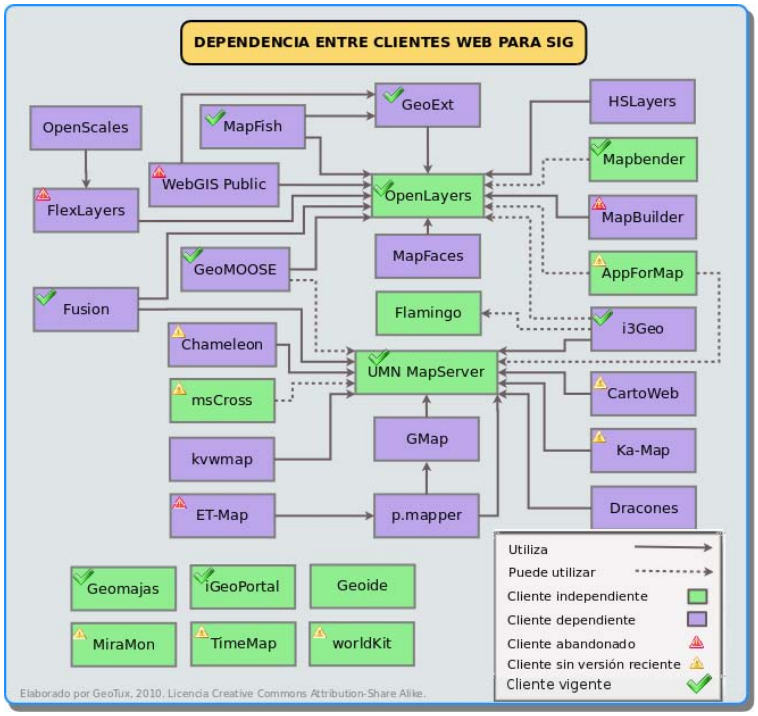


Fig. 3. Dependencia entre clientes web para GIS (Figura obtenida de OSGeo) junto con su vigencia

La mayoría de proyectos mostrados en la Figura 3 giran en torno a dos paradigmas: *UMN MapServer*[16] y *OpenLayers*. Los clientes que utilizan como base UMN MapServer, fueron creados aprovechando las características que este cliente dispone como mapa, escala, mapa de referencia, herramientas de navegación básica, identificación de objetos espaciales y su Interface de Programación de Aplicaciones (API) llamada MapScript[17] que ha sido implementada en diferentes lenguajes de programación. A su vez, algunos clientes utilizan opcionalmente UMN MapServer por medio de MapScript como AppForMap[18] y GeoMOOSE[19]. Por otra parte, la nueva generación de clientes utiliza *OpenLayers* debido a su óptimo rendimiento en las tareas de mostrar los mapas a través de una interface web. Diferentes empresas contribuyen a su

[15] http://wiki.osgeo.org/wiki/Comparacion_de_clientes_ligeros_web_para_SIG/
[16] http://mapserver.org/
[17] http://mapserver.org/mapscript/index.html
[18] http://appformap.mapuse.net/
[19] http://www.geomoose.org/

desarrollo y proyectos como MapBuilder[20] han finalizado para acelerar su progreso. Existen algunos clientes que permiten también elegir una manera adicional para renderizar sus mapas con este paradigma como i3Geo[21] y Flamingo[22]. Por último, existen clientes que no se han basado en otros sino que han sido originados de manera independiente, como *Geomajas*[23], *iGeoPortal*[24], y *Mapbender*[25].

Como la mayoría de estas APIs (Application Programming Interface) para clientes ligeros están escritas en Javascript, en una primera aproximación se buscó encapsular el código, utilizando OpenLayers y GeoExt en servlets. Sin embargo, esta modalidad seguía teniendo los inconvenientes propios del lenguaje (no está totalmente orientado a objetos y no es un lenguaje tipado). Por tal motivo, se descartaron los clientes ligeros cuyo lenguaje era JavaScript, y analizaron diferentes frameworks de desarrollo. Así se llegó al framework desarrollado por Google, Google Web Toolkit[26] (GWT), sobre el que está desarrollado el proyecto de GeoMajas. GWT es una herramienta que facilita la creación de componentes web Javascript. Como se describió anteriormente, el desarrollo de componentes Javascript suele resultar en un proceso tedioso, ya que cada navegador tiene sus propias particularidades y el programador de las aplicaciones debe comprobar que funcionan correctamente en cada uno de los ellos. Por ello GWT proporciona una herramienta que permite al desarrollador programar sus aplicaciones web en lenguaje Java, y que posteriormente al compilarlas devuelven el código Javascript y HTML equivalente. Además, es el propio GWT el encargado de hacer que el Javascript generado funcione correctamente en los distintos navegadores. A partir de la lectura de la documentación de este framework, se encontraron numerosos proyectos vinculados a GWT, entre ellos uno relacionado a OpenLayers, denominado *GWT-OpenLayer*[27]. Este es un wrapper Java para la API JavaScript OpenLayers que permite a los proyectos GWT usarlo como una librería (.jar). Es así que después de numerosas pruebas y análisis, se pudo obtener finalmente la combinación de herramientas que cumplía con los requisitos planteados para implementar la capa de interface de usuario. Por un lado se continúa utilizando OpenLayers, seleccionado por su facilidad, eficiencia y uso, y por otro lado, se pueden desarrollar componentes web utilizando el lenguaje Java. Otra ventaja importante en la elección de esta herramienta fue que GWT soporta invocación de métodos remotos (RPC) permitiendo una perfecta integración con la tecnología EJB.

- *Análisis y Diseño del Suddominio:* La información obtenida en la actividad anterior se utiliza para analizar y organizar las funciones o servicios que el subdominio debe ofrecer junto con las herramientas disponibles para llevarlos a cabo. El modelado del dominio permite encontrar los aspectos comunes y las variabilidades que caracterizan a las aplicaciones dentro del mismo. Como el dominio a analizar se refiere específicamente al dominio geográfico, se deben aplicar las guías definidas en la norma 19109[28] mediante el uso de los tipos espaciales definidos en la norma 19107[29].

  En esta actividad se especificó el modelo conceptual final de acuerdo a los requerimientos nuevos y anteriores obtenidos en las etapas previas. Dicho modelo fue realizado mediante los lineamientos de normas explicadas previamente.
- *Inventario y Análisis de Componentes:* En esta actividad se identifican los componentes directamente de la plataforma heredada que pueden ser extraídos para incluirse como parte de la línea y conformar la arquitectura. Luego como parte del análisis de los componentes candidatos, se seleccionan aquellos tal cual se encuentran implementados y posteriormente se determinan los cambios necesarios para adaptarlos a las nuevas necesidades.

  Aquí se realizó un inventario de los módulos creados en los trabajos previos [14] verificando que

[20] www.mapbuilder.net
[21] http://www.gvsig.org/web/projects/i3Geo
[22] http://www.flamingo-mc.org
[23] http://www.geomajas.org
[24] http://wiki.deegree.org/deegreeWiki/iGeoPortal
[25] http://www.mapbender.org/
[26] http://code.google.com/intl/es-ES/webtoolkit/
[27] http://www.gwt-openlayers.org/
[28] Geographic information. Rules for Application Schema. Draft International Standard 19109, ISO/IEC, 2005
[29] Geographic information. Spatial Schema. International standard 19107, ISO/IEC, 2003

su especificación coincida con la descripción de los servicios que cada uno debía proveer. En dichos trabajos, se utilizó la norma ISO 19119 para derivar los servicios que se debían ofrecer dentro del dominio de ecología marina. En la Tabla 1 se muestra una lista simplificada de los módulos con la información de los servicios provistos por cada uno, según la LPS heredada.

Módulos	**Servicios que implementa**
Interface Gráfica	Mostrar/ocultar capas - Herramientas de zoom - Herramientas de desplazamiento, Escala
Administrador de Características Geográficas	Mostrar y consultar datos sobre características geográficas - Ver, consultar y editar datos de características geográficas gráficamente
Detección de Cambios	Encontrar diferencias en los datos de un mismo tipo dentro de un área geográfica específica en distintos momentos
Análisis de Proximidad	Obtener todas las características geográficas dentro de un área específica
Estadísticas Gráficas	Generar estadísticas con datos de características gráficas
Acceso a Características Geográficas	Realizar consultas a un repositorio de características geográficas - Administrar los datos de las características geográficas
Acceso a Mapas	Acceder a mapas geográficos

Table 1. Inventario de módulos

A partir de este inventario de módulos y de los nuevos requerimientos solicitados se realizó un análisis para obtener los componentes candidatos. Para esto se utilizó la lista de servicios provista en los trabajos previos (derivada de la ISO 19119 para el dominio de ecología marina) adicionándole los servicios necesarios para cumplir con los nuevos requerimientos. Posteriormente se creó la nueva arquitectura de referencia para la LPS basada en las tres capas (Figura 2). En la Tabla 2 se muestran los componentes candidatos junto a la descripción de los servicios que proveen, agrupados de acuerdo a las capas mencionadas anteriormente.

Componentes	**Descripción**	**Capa de Arquitectura**
Visor Geográfico	Muestra el espacio geográfico total - Muestra/oculta capas	Interface de Usuario
Visor de Atributos	Muestra datos sobre características geográficas - Muestra detalles de las capas existentes - Muestra gráficos adicionales	Interface de Usuario
Características de Visualización	Funcionalidad de zoom - Desplazamiento - Escala - Refresco	Interface de Usuario
Herramientas de Manipulación Geográfica	Permite realizar figuras geométricas sobre el mapa, tales como puntos, líneas polígono	Interface de Usuario
Interfaz Gráfica	Permite ingresar datos - Permite visualizar los resultados de las distintas operaciones realizadas	Interface de Usuario
Análisis de Proximidad	Obtiene todas las características geográficas dentro de un área específica	Procesamiento Geográfico
Detección de Cambios	Encuentra diferencias en los datos de un mismo tipo dentro de un área geográfica específica en distintos momentos	Procesamiento Geográfico
Estadísticas Gráficas	Permite generar distintos gráficos estadísticos a partir de datos de diversas características	Procesamiento Geográfico
Consultar datos de la Base de Datos	Realiza distintos tipos de consulta acerca de la información almacenada	Procesamiento Geográfico
Acceso a Características Geográficas	Realiza consultas a un repositorio de características geográficas - Administrar los datos de las características geográficas	Modelado Geográfico
Acceso a Mapas	Accede a mapas geográficos	Modelado Geográfico

Table 2. Detalle de componentes de la nueva LPS

En la capa *Interface de Usuario* el módulo *Administrador de características geográficas* poseía tanto los servicios de visualizador de datos geográficos como de editor de características geográficas. Para mejorar la modificabilidad y evolución, se decidió crear tres nuevos componentes separando dichos servicios con las siguientes funcionalidades: el componente *Visor geográfico*, cuya funcionalidad es mostrar el espacio geográfico total y mostrar/ocultar capas; el componente *Visor de atributos* que muestra un detalle de las capas existentes, permite seleccionar y de-seleccionar capas y muestra gráficos adicionales; y el componente *Características de visualización* que posee las funcionalidades de zoom, desplazamiento, escala y refresco. Se agregó también el componente *Herramienta de manipulación gráfica* que permite realizar figu-

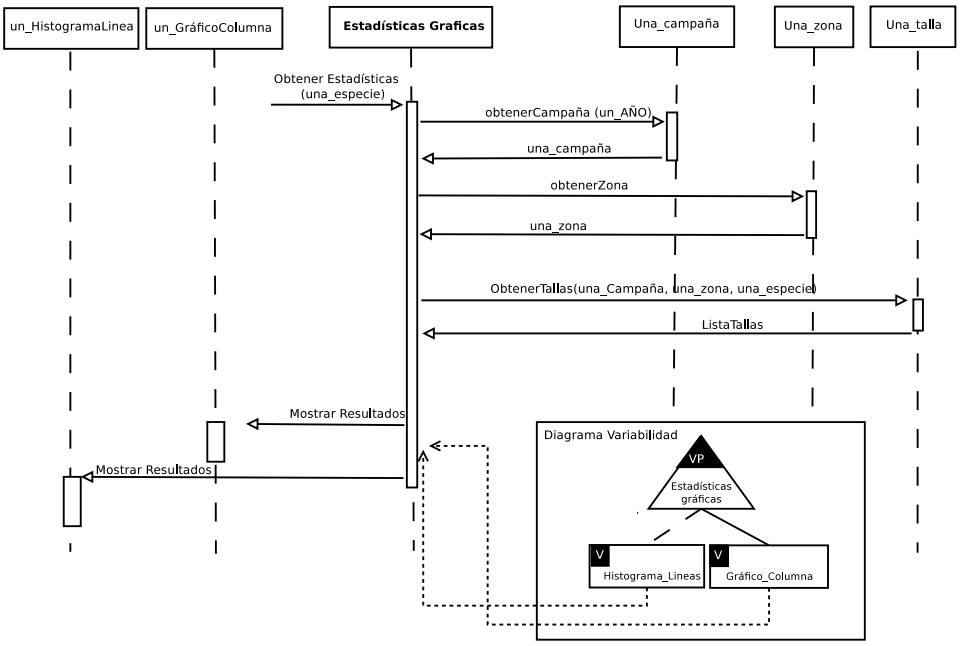


Fig. 4. Diagrama de secuencia correspondiente al componente *Estadísticas Gráficas*

ras geométricas, tales como puntos, líneas y polígonos, y luego enviar la información para su análisis.

En la capa de *Procesamiento geográfico* se realizaron varias modificaciones y se agregaron nuevos componentes de acuerdo a los requerimientos. Por ejemplo, el componente de *Análisis de proximidad* se mantuvo con la misma funcionalidad que poseía el módulo del mismo nombre, al igual que *Detección de cambios*. El componente *Estadísticas gráficas* en cambio, se redefinió completamente, permitiendo obtener y analizar estadísticas a través de los datos obtenidos del componente *Consultar datos de la base de datos*. Los componentes *Exportar mapas* y *Conteo de características geográficas* se desarrollarán en trabajos futuros. Por último en la capa de *Modelado geográfico* se crearon componentes con la misma funcionalidad que los módulos que se encontraban previamente, pero se reimplementaron con las nuevas herramientas.

– *Análisis y Diseño de la Plataforma:* En esta actividad se crea la arquitectura de referencia final basada en las características definidas en los procesos anteriores y la estructura de componentes. Se plantea la forma de extracción y desarrollo de componentes y se toman las decisiones sobre la asignación de la funcionalidad y la variabilidad de los mismos.
En nuestro trabajo se creó la arquitectura de referencia final de la LPS basada en los componentes y las capas definidas. A su vez, debido a los nuevos requerimientos surgidos del proceso de reestructuración, se diseñaron y desarrollaron dos nuevos componentes, *Cálculos Mapa* y *Estadísticas Gráficas* que implementan servicios para la ejecución de determinadas consultas espaciales y la generación de datos estadísticos. Este último componente, permite mostrar determinados datos de las especies en forma de histograma, a fin de analizar y comparar los resultados de la recolección de muestras en las diferentes zonas y campañas. La Figura 4 muestra el diagrama de secuencia que modela la interacción entre los objetos.
Como podemos observar, a partir de una campaña, un año, una zona y una especie se obtienen los datos de las tallas de los individuos (medida en mm). Con esta información se muestran los datos en el histograma. En este componente se observa un punto de variabilidad ya que los resultados se pueden mostrar a través de *histogramas en forma de líneas continuas* o bien en *gráficos de barras*.
– *Implementación de la Plataforma:* En esta actividad se implementan todos aquellos componentes de la plataforma de la LPS.
En nuestro trabajo de reestructuración se reimplementaron todos los componentes según la funcionalidad definida en la Tabla 2 y los requerimientos de la arquitectura de referencia. Por

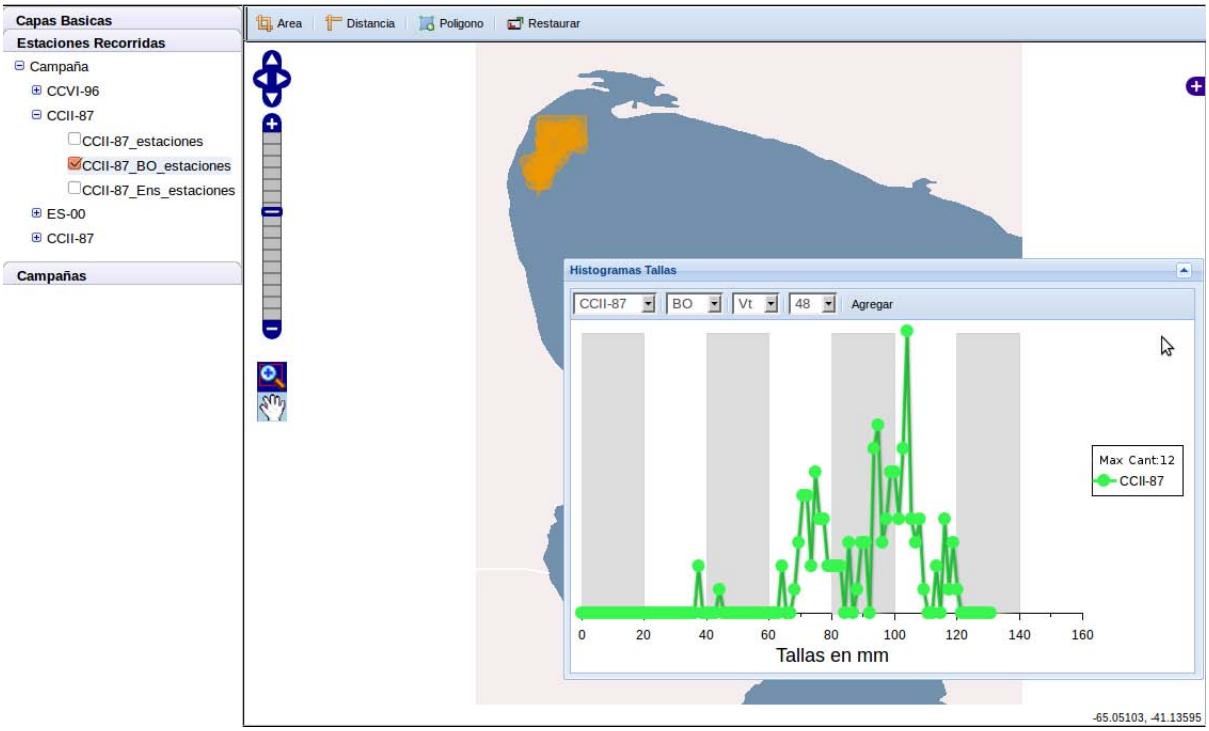


Fig. 5. Pantalla mostrando la interface del servicio *Estadísticas Gráficas*

ejemplo, la Figura 5 muestra la interface de usuario creada para el componente de *Estadísticas gráficas* (ver Figura 4) a través de histogramas en forma de líneas continuas.
El principal objetivo del servicio de estadísticas es lograr una visualización sencilla y rápida del rango de las tallas en las que se encontraron la mayor cantidad de individuos. Por ejemplo, en la figura podemos ver que para la campaña *CCII 87* de la *Viera Tehuelche* la mayor cantidad de individuos encontrados miden entre 60 y 120 mm.

– *Validación:* Se aplican casos de prueba para verificar la plataforma y la especificación de la línea de productos. En esta etapa también se deben realizar las pruebas de los nuevos productos derivados de la línea desarrollada.
Se han realizado pruebas analizando la funcionalidad sobre la nueva implementación de la LPS, en especial sobre los dos productos derivados de la misma. La LPS ya definida como componentes reusables fue utilizada para instanciar dos nuevos productos, uno para el IBMPAS y otro para el Centro Nacional Patagónico[30] (CENPAT-CONICET). Dichos productos están disponibles en `http://gissrv.fi.uncoma.edu.ar/SaoProjectUI` y `http://gissrv.fi.uncoma.edu.ar/CenpatProjectUI` respectivamente.

3 Conclusión y Trabajo Futuro

En este artículo se describió la reestructuración de una línea de productos de software a un enfoque orientado a componentes reusables, creados según las particularidades del dominio geográfico, y se lo ilustró de acuerdo a un caso de estudio. La metodología utilizada se compone de una serie de actividades que incluyen la utilización de estándares geográficos y de previas clasificaciones de servicios en el subdominio de ecología marina. El uso de dichos estándares permite delimitar el rango de servicios que el dominio debería ofrecer posibilitando que sea ampliado a otros subdominios geográficos diferentes al de ecología marina. Así la metodología permite reusar también los servicios definidos minimizando el tiempo y esfuerzo en las primeras etapas de la misma. Luego, el caso de estudio descripto muestra varios beneficios derivados de la utilización de dichos estándares, tanto en la especificación de los componentes como en su influencia en la selección de las herramientas

[30] http://www.cenpat.edu.ar/

de código abierto para la implementación de los mismos. En este trabajo la selección de dichas herramientas fue una de las tareas más complejas y que demandó más tiempo debido a que no sólo había que contemplar que las mismas permitan el desarrollo de componentes según la arquitectura definida en la norma ISO 19119, sino que también posean continuidad en sus desarrollos, buenas documentaciones, foros de consulta activos, flexibilidad para ser extendidos, etc. El resultado final del trabajo se traduce en un conjunto de componentes altamente cohesivos y débilmente acoplados que maximizan la modificabilidad, evolución y especialmente el reuso, ya que permiten ser tratados como unidades independientes para ser ensambladas luego en la creación de la plataforma de una LPS. Así, los tiempos y el esfuerzo dedicados en la creación de nuevos productos se pueden ver disminuídos debido a la simplicidad en el uso y ensamblaje de los componentes de la línea. Como trabajos futuros, se debe validar la nueva metodología aplicándola a otros proyectos de reingeniería de LPS. A su vez, se deben medir los logros alcanzados mediante indicadores de reuso que permitan mostrar a nivel cualitativo y cuantitativo del reuso alcanzado tanto en la creación de una LPS basada en componentes como en la creación de los nuevos productos derivados de ella.

References

1. F. Amin, A. K. Mahmood, and A. Oxley. Reusability Assessment of Open Source Components for Software Product Lines. *International Journal of New Computer Architectures and their Applications (IJNCAA)*, 3(1), 2011.
2. C. Atkinson, J. Bayer, and D. Muthig. Component-based product line development: The kobra approach. 576:289–309, 2000.
3. J. Bosch. *Design and use of software architectures: adopting and evolving a product-line approach*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 2000.
4. Agustina Buccella, Alejandra Cechich, Maximiliano Arias, Matias Pol'la, Maria del Socorro Doldan, and Enrique Morsan. Towards systematic software reuse of gis: Insights from a case study. *Computers & Geosciences*, 54(0):9 – 20, 2013.
5. P. C. Clements and L. Northrop. *Software Product Lines : Practices and Patterns*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.
6. K. Czarnecki, S. Helsen, and U. W. Eisenecker. Formalizing cardinality-based feature models and their specialization. *Software Process: Improvement and Practice*, 10(1):7–29, 2005.
7. G. T. Heineman and W. T. Councill, editors. *Component-based software engineering: putting the pieces together*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.
8. Natalia Huenchuman. Reestructuración de una línea de productos de software para el subdominio de ecología marina. Tesis de Licenciatura en Ciencias de la Computación, 2013.
9. Waraporn Jirapanthong. Experience on re-engineering applying with software product line. *CoRR*, abs/1206.4120, 2012.
10. K. Kang, S. Cohen, J. Hess, W. Nowak, and S. Peterson. Feature-Oriented Domain Analysis (FODA) Feasibility Study. Technical Report CMU/SEI-90-TR-21, Software Engineering Institute, Carnegie Mellon University Pittsburgh, PA., 1990.
11. M. Laguna and Y. Crespo. A systematic mapping study on software product line evolution: From legacy system reengineering to product line refactoring. *Science of Computer Programming*, 78(8):1010–1034, 2013.
12. D. Mcilroy. Mass-produced Software Components. In *Proceedings of Software Engineering Concepts and Techniques*, pages 138–155. NATO Science Committee, January 1969.
13. P. Pernich, A. Buccella, A. Cechich, S. Doldan, and E. Morsan. Reusing geographic e-services: A case study in the marine ecological domain. In Wojciech Cellary and Elsa Estevez, editors, *Software Services for e-World*, volume 341 of *IFIP Advances in Information and Communication Technology*, pages 193–204. Springer Boston, 2010.
14. P. Pernich, A. Buccella, A. Cechich, S. Doldan, E. Morsan, M. Arias, and M. Pol'la. Product-line instantiation guided by subdomain characterization: A case study. *Journal of Computer Science and Technology, Special Issue 12(3). ISSN: .*, 12(3):116–122, 2012.
15. Klaus Pohl, Günter Böckle, and Frank J. van der Linden. *Software Product Line Engineering: Foundations, Principles and Techniques*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
16. D. Smith, O. Liam, and J. Bergey. Using the options analysis for reengineering (oar) method for mining components for a product line. 2379:316–327, 2002.
17. Clemens A. Szyperski. *Component software - beyond object-oriented programming*. Addison-Wesley-Longman, 1998.
18. Frank van der Linden, Klaus Schmid, and Eelco Rommes. *Software Product Lines in Action: The Best Industrial Practice in Product Line Engineering*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.
19. G. Zhang, L. Shen, X. Peng, Z. Xing, and W. Zhao. Incremental and iterative reengineering towards software product line: An industrial case study. In *Proceedings of the 2011 27th IEEE International Conference on Software Maintenance*, ICSM '11, pages 418–427, Washington, DC, USA, 2011. IEEE Computer Society.

Generación Automática del Modelo de Diseño desde el Modelo de Análisis a través de Reglas QVT

Ariel Arsaute, Fabio Zorzan, Marcela Daniele, Paola Martellotto

Dpto. de Computación, Facultad de Ciencias Exactas, Fco-Qcas y Naturales
Universidad Nacional de Río Cuarto
Río Cuarto, Córdoba, Argentina
{aarsaute, fzorzan, marcela, paola}@dc.exa.unrc.edu.ar

Abstract. Model Driver Architecture (MDA) define un proceso de construcción del software basado en producción y transformación de modelos. MDA se fundamenta en los principios de abstracción, automatización y estandarización. Vinculado con MDA, la Object Management Group (OMG) ha definido el estándar Query/View/Transformation (QVT) para la definición y transformación de modelos de software. Por otro lado, el Proceso Unificado (PU), también define un proceso de construcción del software generando distintas vistas o modelos. En este trabajo se sientan las bases para la integración de MDA y el PU. Se propone un conjunto de reglas QVT que establecen una transformación de forma automática entre los modelos producidos en las etapas de Captura de Requisitos, Análisis Y Diseño. El objetivo del trabajo es la definición del conjunto de reglas QVT que posibiliten la transición desde la etapa de Análisis a la etapa de Diseño considerando la tecnología de implementación RemoteMethodInvocation (RMI).

Keywords:ProcesoUnificado, MDA, Relations, QVT.

1 Introducción

La dinámica propia de la Ingeniería de Software implica el surgimiento constante de nuevas tecnologías que den soporte al proceso de construcción de sistemas de información. Todo esto implica un arduo trabajo de integración que posibilite la co-existencia de las tecnologías involucradas.

MDA [1] define un proceso de construcción de software basado en la producción y transformación de modelos. MDA se fundamenta en los principios de abstracción, automatización y estandarización. La idea principal subyacente en MDA es abstraer propiedades y características de los sistemas de información en un modelo abstracto independiente de los cambios producidos en las tecnologías.

En sintonía con MDA, la OMG, ha definido el estándar QVT [2]. QVT consta de consultas, vistas, y reglas de transformación. El componente de consultas de QVT recibe como entrada un modelo, y selecciona elementos específicos. Una vista puede verse como el resultado de una consulta sobre un modelo que proporciona un punto

de vista particular. El componente de transformaciones de QVT permite definir transformaciones entre modelos. Dichas transformaciones describen relaciones entre dos metamodelos: fuente y objetivo. Ambos metamodelos deben ser especificados en Meta ObjectFacility (MOF) [3]. Una vez definida y aplicada la transformación, se obtiene el modelo instancia del metamodelo objetivo, a partir de un modelo fuente.

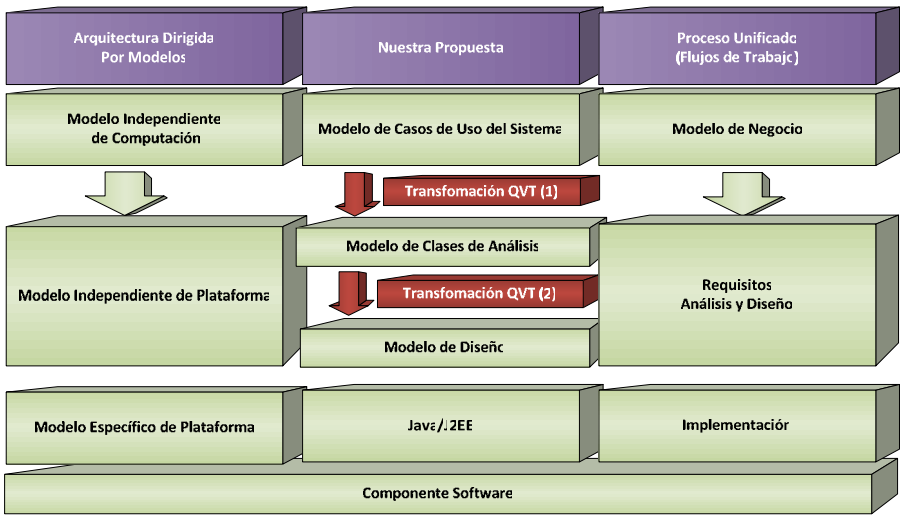


Fig. 1. Propuesta de Transformación.

Por otro lado, el Proceso Unificado [4] es una metodología de desarrollo de software que define tareas y responsabilidades para la construcción de un producto de software. La metodología divide el desarrollo en etapas, cada una de las cuales produce diferentes modelos o vistas del sistema. Las primeras etapas del proceso, centran sus esfuerzos en la comprensión del problema, las tecnologías a utilizar, la delimitación del ambiente del proyecto, el análisis de los riesgos, y la definición de la arquitectura del proyecto. Las actividades centrales son aquellas encargadas de modelar el negocio.

El objetivo final de esta línea de investigación, es lograr la automatización parcial de las actividades del PU, desde la Captura de Requisitos hasta la Implementación, aplicando reglas de transformación de modelos QVT. En este artículo se presenta el segundo paso para lograr este objetivo, que corresponde a la Transformación QVT (2) expresada en la Fig.1. Esta transformación contiene las reglas QVT que permitirán pasar del modelo de análisis al modelo de diseño. El modelo de análisis, nuestro modelo fuente, corresponde a una instancia del metamodelo de UnifiedModelingLanguage (UML) [5], el cual se obtuvo como resultado de aplicar la Transformación QVT (1) definida en [6] por los autores de este trabajo, y el modelo de diseño, nuestro modelo objetivo, también corresponde a una instancia del metamodelo UML. Como resultado, la aplicación de las reglas QVT producen un diagrama de clases de diseño (etapa de diseño) a partir de un diagrama de clases de análisis (etapa de análisis).

El artículo está organizado de la siguiente forma: En la sección 2 se presenta MDA, en la sección 3 se referencia al estándar QVT definido por la OMG. En la sección 4 se presenta la metodología de desarrollo del Proceso Unificado [4] como también tecnología y patrones que tuvimos en cuenta en el diseño. En la sección 5 se presenta la propuesta junto con un ejemplo de aplicación, y finalmente, en la sección 6 se expone las conclusiones.

2 Model Driver Architecture (MDA)

MDA define un proceso de construcción de software basado en la producción y transformación de modelos. Los principios en los cuales se fundamenta MDA son: abstracción, automatización y estandarización. El proceso central de MDA es la transformación de modelos. La idea principal subyacente es utilizar modelos, de manera que las propiedades y características de los sistemas queden contenidas en un modelo abstracto independiente de los cambios producidos en la tecnología. MDA proporciona una serie de guías o patrones expresadas como modelos.

Con respecto a la automatización, MDA favoreció al surgimiento de nuevas herramientas CASE con funcionalidades específicas destinadas al intercambio de modelos, verificación de consistencia, transformación de modelos y manejo de metamodelos, entre otras.

3 Query/View/Transformation (QVT)

La OMG [7] ha definido el estándar QVT [2] para trabajar con modelos de software. QVT consta de consultas, vistas, y transformaciones. El componente de consultas de QVT toma como entrada un modelo, y selecciona elementos específicos del mismo. Para la resolución de las consultas, se propone el uso de una versión extendida de OCL 2.0 [8]. Una vista es una proyección realizada sobre un modelo, creada mediante una transformación. Una vista puede verse como el resultado de una consulta sobre un modelo. En esta sección se presenta el componente de transformaciones de QVT que tiene como objetivo definir transformaciones entre modelos. Estas transformaciones describen relaciones entre un metamodelo fuente F y un metamodelo objetivo O, ambos metamodelos deben ser especificados en MOF [3]. Una vez definida la transformación, es utilizada para obtener un modelo objetivo, una instancia del metamodelo O, a partir de un modelo fuente, que es una instancia del metamodelo F. También la transformación puede ser utilizada para chequear la correspondencia entre dos modelos. La especificación de QVT 1.1 [2] tiene una naturaleza híbrida declarativa/imperativa. En este trabajo interesa el lenguaje Relations que tiene naturaleza declarativa.

3.1 Lenguaje Relations

El lenguaje Relations es una especificación declarativa de las relaciones entre metamodelos MOF. Una transformación especifica un conjunto de relaciones que deben cumplir los elementos de los modelos involucrados. La ejecución de una transformación en una dirección particular se realiza seleccionando el metamodelo objetivo de la transformación. Una relación especifica la relación entre elementos de los modelos candidatos. La relación involucra dos o más dominios, y dos restricciones denominadas cláusulas guard (o cláusula when) y cláusula where. La cláusula guard de la relación especifica la condición bajo la cual la relación debe ser cumplida. La cláusula where define la condición que deben cumplir los elementos intervinientes en la relación. Una relación puede ser declarada en modo sólo chequeo (checkonly) o forzado (enforced). Si un domino es marcado como sólo chequeo, cuando se ejecute la transformación sólo será chequeado para ver si existe una correspondencia válida con otro dominio perteneciente al modelo objetivo; en cambio, si el modelo está marcado como forzado, cuando una relación no se satisface, los elementos del modelo objetivo serán creados, borrados o eliminados en el modelo para satisfacer la relación. En la actualidad existen herramientas bastantes maduras que implementan el lenguaje Relations, una de estas es MediniQVT[9].

3.2 MediniQVT

Esta herramienta implementa la especificación QVT/Relations de la OMG en un poderoso motor QVT. Está diseñada para transformaciones de modelos permitiendo un rápido desarrollo, mantenimiento y particularización de reglas de transformación de procesos específicos. La herramienta está integrada a Eclipse y utiliza EMF para la representación de modelos. Además, posee un editor con asistente de código y un depurador de relaciones.

4 Proceso Unificado: Del Análisis al Diseño

La metodología del Proceso Unificado [4]divide el desarrollo en etapas, cada una de las cuales produce diferentes modelos o vistas del sistema. Las primeras etapas del proceso, centran sus esfuerzos en la comprensión del problema, las tecnologías a utilizar, la delimitación del ambiente del proyecto, el análisis de los riesgos, y la definición de la arquitectura del proyecto. Para la arquitectura del proyecto eneste trabajo se asume las siguientes consideraciones arquitecturales que permiten definir el diseño a generar: Arquitectura distribuida utilizando la tecnología Java RMI [10], uso de patrón de Arquitectura Layer [11], patrones de diseño Mediador [12] y DTO [13].

4.1 Remote Method Invocation (RMI)

RemoteMethodInvocation (RMI) [10] es un mecanismo que permite realizar llamadas a métodos de objetos remotos situados en distintas (o la misma) máquinas

virtuales de Java, compartiendo así recursos y carga de procesamiento a través de varios sistemas.

Toda aplicación RMI normalmente se descompone en 2 partes:

- **Un servidor**, que crea algunos objetos remotos, crea referencias para hacerlos accesibles, y espera a que el cliente los invoque.
- **Un cliente**, que obtiene una referencia a objetos remotos en el servidor, y los invoca.

Las aplicaciones Java que utilizan RMI deben contener básicamente del lado del servidor una clase de control. Esta clasedebe ofrecer métodos a ser invocados remotamente por un cliente, estas clases de control deben implementar Interfaces remotas. Las interfaces remotas son las utilizadas para la comunicación entre el cliente y servidor. RMI permite desarrollar aplicaciones distribuidas muy fácilmente. Entre otras ventajas de utilizar de RMI en aplicaciones Java se pueden nombrar: permite una separación entre la interface y la implementación, es posible la descarga dinámica de código y permite utilizar protocolos seguros de comunicación como SSL y HTTPS.

4.2 Data Transfer Object (DTO)

Data Transfer Object (DTO) es un patrón de diseño [11] que ofrece una solución para distribución de objetos, en particular para el intercambio de datos en llamadas remotas costosas.

Cuando se trabaja con una interfaz remota, como RMI, cada llamada remota es costosa. Como resultado, usted necesita reducir el número de llamadas, y eso significa que usted necesita para cada llamada transferir más datos.

La solución es crear un objeto de transferencia de datos (DTO) que puede contener todos los datos de la llamada. Un DTO es un objeto serializable que puede viajar fácilmente a través de la red y que contiene generalmente información perteneciente a varios objetos de dominio. Por lo general, se utiliza un ensamblador en el lado del servidor para transferir datos entre el DTO y los objetos de dominio.

El patrón Data Transfer Object [13] permite crear el objeto que será serializado para transferir información entre la interfaz del usuario (frontend) y el servidor (back end) del sistema.De esta manera se consigue un objeto serializable para pasar entre las dos capas, con la información necesaria para reducir el número de llamadas entre ellas.

5 Automatización del mapeo entre clases deanálisis a clases de diseño

El objetivo final de esta línea de investigación es lograr la automatización parcial de las actividades del PU, es decir, desde la “Captura de Requisitos” hasta la “Implementación”, aplicando reglas de transformación de modelos QVT. En este trabajo se presenta la segunda etapa que consiste en la transformación del modelo de

análisis al modelo de diseño como continuidad del trabajo previo [6], donde se realizó la transformación del modelo de CU de sistema al modelo de análisis. Como se mencionó anteriormente, una transformación en QVT requiere al menos dos modelos, uno fuente y otro objetivo. En este caso, el fuente es una instancia del metamodelo UML correspondiente al modelo de análisis y el objetivo también corresponde a una instancia del metamodelo UML, el cual es el modelo de diseño. Los modelos genéricos de análisis y diseño se muestran en las Fig. 2 y 3respectivamente.

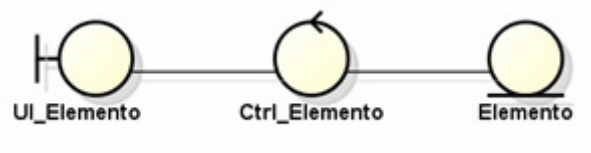


Fig. 2. Modelo genérico de análisis fuente de la transformación

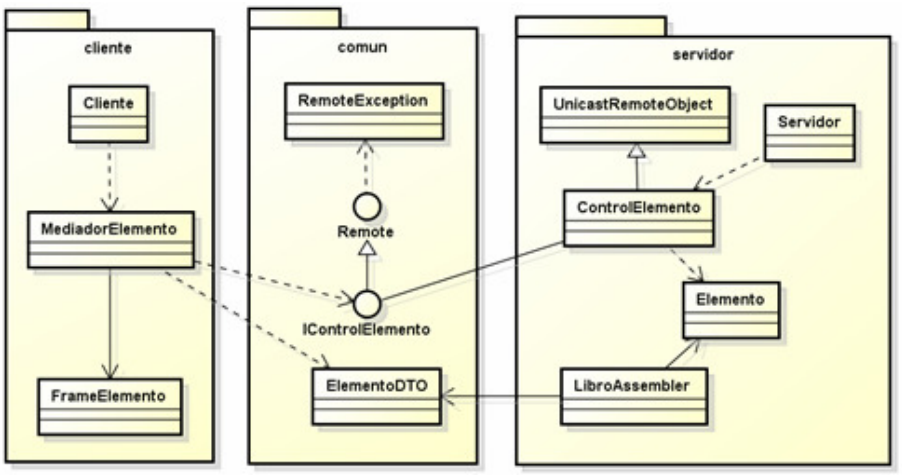


Fig. 3. Modelo genérico de diseño generado por la aplicación de la transformación

La herramienta CASE utilizada para la definición de las reglas QVT es MediniQVT [9]. Esta herramienta fue presentada en la seccion 3.2 de este trabajo.

La Fig. 4 muestra, en forma abreviada, la sintaxis gráfica de la relación packageAnalisisToPackageDisenio, esta relación es la más importante de la transformación entre los metamodelos UML de análisis y diseño

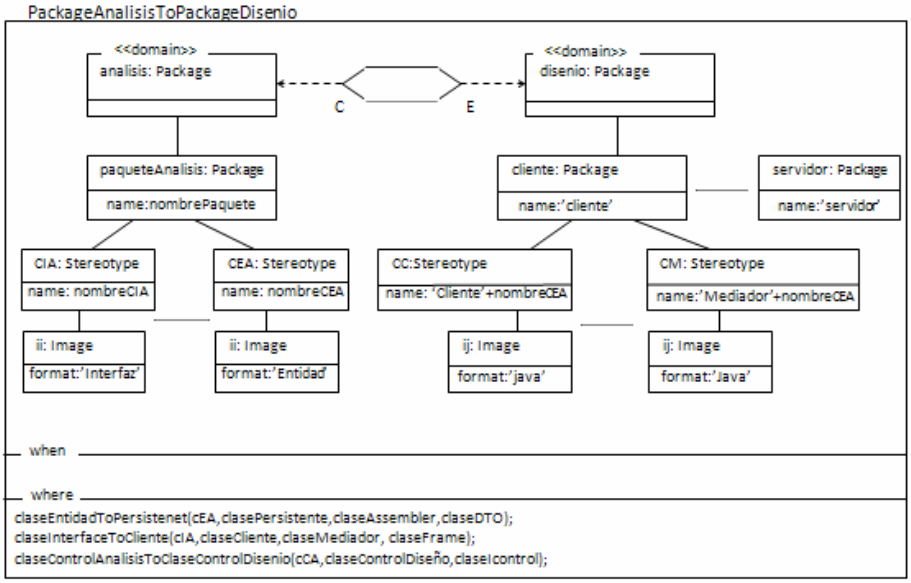


Fig. 4. Sintaxis Gráfica de laRelación

La transformación QVT se define a nivel Metamodelos y se aplica a nivel modelos. La transformación se define en Relations de la siguiente manera:

```
transformationAnalisisToDisenio
(modeloAnalisis:uml ,modeloDisenio:uml)
```

Esta transformación toma un modelo de análisis (modeloAnalisis), que es una instancia del metamodelo UML y un modelo de diseño (modeloDisenio) que también es una instancia del metamodelo UML.

A continuación se muestra la definición en Relations de la relación más importante de la transformación la cual se denominaPackageAnalisisToPackageDisenio.Cabe aclarar que en la definición presentada sólo se expone en forma completa la obtención del paquete *cliente*del diseño generado, por una cuestión de tamaño de la relación, sólo se muestra una parte de la definición de los paquetes *comun*y *servidor*.

```
toprelationPackageAnalisisToPackageDisenio {
nombrePaquete : String;
checkonlydomainmodeloAnalisisanalisis : uml::Package {
packagedElement =paquete_analisis:uml::Package
{name=nombrePaquete,
 // cIA clase Interfaz de Analisis
packagedElement = cIA:uml::Stereotype
     {name= nombreClaseInterfaz,
     icon = ii :uml::Image {format = 'Interfaz'}},
 //cCAclase Control de Analisis
```

```
packagedElement = cCA : uml::Stereotype
        {name= nombreClaseControl,
     icon = ic :uml::Image {format ='Control'}},
//cEAclase Entidad de Analisis
packagedElement = cEA: uml::Stereotype
        {name= nombreClaseEntidad,
     icon = ie :uml::Image {format = 'Entidad'}},
name = nombrePaquete
   };
enforcedomainmodeloDiseniodisenio: uml::Package {
packagedElement =cliente : uml::Package{name='cliente',
//cC clase control
packagedElement = cC : uml::Stereotype
{name='Cliente',
icon = ij :uml::Image {format = 'Java'}},
//cM clase Mediador
packagedElement = cM : uml::Stereotype
 {name='Mediador'+nombreClaseEntidad,
icon = ij :uml::Image {format = 'Java'}},
//cFclase Frame
packagedElement = cF : uml::Stereotype
{name='Frame'+nombreClaseEntidad,
icon = ij :uml::Image {format = 'Java'}},
packagedElement =comun : uml::Package
 {name='comun',
       packagedElement =claseRemoteException:uml::Stereotype
              :
              :     },
packagedElement =servidor:uml::Package
     {name='servidor'
:  },
name = nombrePaquete
};
where {
claseEntidadToPersistente(cEA,clasePersistente,
claseAssembler,claseDTO);
claseInterfaceToCliente(cIA,claseCliente,claseMediador,
claseFrame);
claseControlAnalisisToClaseControlDisenio(cCA,
claseControlDiseño,claseIcontrol);
}
}
```

La relación PackageAnalisisToPackageDisenio define la transformacióndel paquete UML del modelo de análisis al modelo de diseño.

En la especificación de la relación se define la transformación de cada paquete de análisis a su correspondiente paquete de diseño, siendo ambos paquetes UML. El paquete de diseño se organiza a su vez en tres paquetes que son:*cliente, común* y *servidor*; que según las consideraciones del diseño y a la experiencia de éste equipo con aplicaciones RMI son los más adecuados

A partir del modelo de análisis con clases de tipo Interfaz, control y entidad, se generan 3 paquetes de diseño, estos son: *cliente*, *comun* y *servidor*. Estos paquetes se generan debido a que la tecnología de implementación será RMI, y como se explicó en la sección 4.1, estos paquetes son necesarios. Dentro del paquete cliente se construirá una única clase Cliente, que es la encargada de iniciar el programa cliente de la aplicación y recuperar todas las interfaces remota para poder interactuar con el servidor. También, se generan en el paquete, por cada clase Interfaz del análisis, una clase Frame y Mediador, la clase Frame es la encargada de interactuar con el usuario, y la clase Mediador es la encargada de mediar entre la Clase Frame y las clases de control del servidor. Dentro del paquete *comun* se generan, por cada clase de tipo control de análisis, una Interfaz Java. Estas interfaces son las que se publicarán por medio del demonio RMI cuando se inicialice el servidor de la aplicación. Además en el paquete *comun* de diseño, por cada clase Entidad de Análisis se genera una clase Java DTO que es la encargada de llevar y traer información del entre el cliente y servidor (ver sección 4.2). Por último, en el paquete *servidor,* se genera una única clase Servidor que es la encargada de iniciar la aplicación servidora publicando las interfaces RMI para que las aplicaciones cliente se puedan conectar. También, se generan, por cada clase de tipo Entidad del análisis, una clase Persistente y otra Assembler que es la encargada de hacer la traducción entre la clase Persistente y la DTO correspondiente del paquete *comun*. Las últimas clases generadas en el paquete servidor son las clases de control, estas clases se generan a partir de cada clase de tipo control del análisis.

Las diferentes relaciones entre las clases son generadas mediante las relaciones `claseEntidadToPersistente`, `claseInterfaceToCliente` y `claseControlAnalisisToClaseControlDisenio`que son parte de la cláusula*where* de la relación.

5.1 Ejemplo de Aplicación

En esta sección se presenta un ejemplo de aplicación de la transformación QVT definidas en el apartado anterior.

En la Fig. 5 se muestra parte del diagrama de clases de análisisdel caso de uso gestión de libros obtenido del ejemplo de aplicación del trabajo previo, y que es modelo fuente de la transformación del presente trabajo. En la Fig.6 se muestra el modelo de análisis en la vista generada por el EcoreModel Editor. En la Fig. 7 se presenta el diagrama de clases de diseño obtenido por aplicación de la transformación al modelo de análisis de la gestión de Libros.

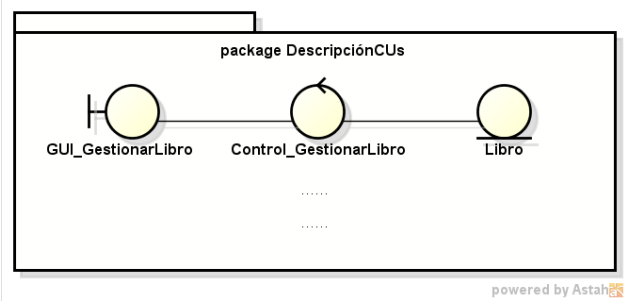


Fig. 5. Diagrama de clases de Análisis - Caso de Uso gestionar

Para llevar a cabo esta transformación se utilizó la herramienta MediniQVT. Teniendo en cuenta que MediniQVT utiliza EMF para representar los modelos/metamodelos involucrados en las transformaciones. Cabe aclarar que el metamodelo UML en formato EMF, que es el metamodelo fuente y objetivo en la definición de la transformación, viene provisto por la herramienta. Para poder aplicar la transformación, que es lo más importante del caso de estudio, fue necesario obtener el modelo fuente, el cual es el resultado de aplicar la transformación realizada en [6]. Luego de obtener los modelos/metamodelos, se aplicó la transformación al modelo fuente y se obtuvo el modelo objetivo. Este modelo corresponde a una especificación UML de diagrama de clases de diseño.En la Fig. 8 se muestra el modelo de diseño en la vista generada por el EcoreModel Editor.

Fig. 6. Diagrama de clases de Análisis - Caso de Uso gestionar Libro, vista generada por el EcoreModel Editor.

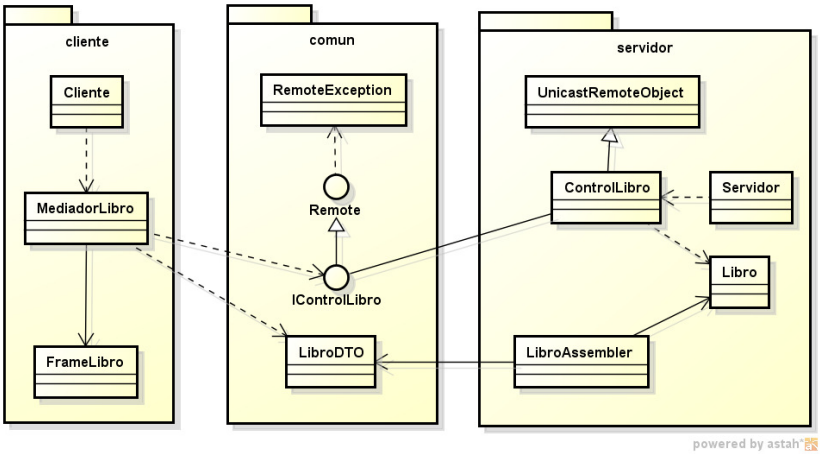


Fig. 7. Diagrama de clases de Diseño - Caso de Uso "Gestionar Libro".

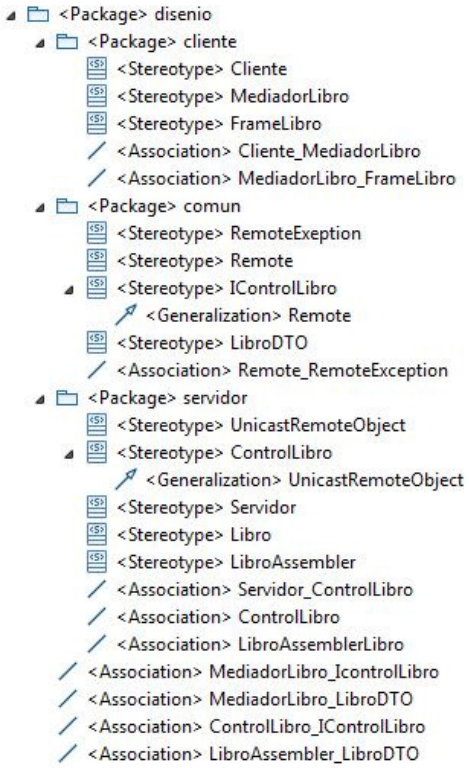


Fig. 8. Diagrama de clases de Diseño obtenido por la aplicación de la transformación, vista generada por el EcoreModel Editor.

6 Conclusiones

Este trabajo tiene como objetivo hacer una contribución a la mejora de los procesos de desarrollo de software. La elaboración y culminación exitosa de este caso de estudio permitió la utilización de una especificación de captura de requerimiento para dar inicio a la construcción del software de manera automática en el marco de MDA utilizando reglas de transformación QVT.

El beneficio de esta transformación se refleja también en el dinamismo de los cambios en los requisitos durante toda la vida del software desarrollado, es decir, cualquier cambio en la especificación de la captura de requerimientos podrá ser propagado automáticamente a los distintos artefactos producidos en el proceso de desarrollo del software, esto debido a que se definió una transformación que puede ser ejecutada con una herramienta permitiendo así adaptar rápidamente los cambios de la captura de requerimiento a la especificación del análisis y diseño.

En esta línea de Investigación, hasta el momento se ha avanzado en la transformación que convierte la especificación de Captura de Requerimientos a una especificación UML correspondiente al artefacto de diseño, pasando por sus respectivas transformaciones. El próximo objetivo es realizar la transformación hacia el modelo final correspondiente al componente de software y de esta manera cubrir con todas las etapas del PU.

Por último, este trabajo intenta promover el uso eficiente de modelos de sistemas en el proceso de desarrollo de software (desarrollo de software dirigido por modelos) que es uno de los principales objetivos de MDA, en el marco de la Ingeniería de Software dirigida por modelos, representando para los desarrolladores, una nueva manera de organizar y administrar arquitecturas empresariales, basada en la utilización de herramientas de automatización de etapas en el ciclo de desarrollo del software. De esta forma, permitir definir los modelos y facilitar trasformaciones paulatinas entre diferentes modelos.

Referencias

1. Miller, J., Mukerji, J., MDA Guide Version 1.0.1 Document number omg/2003-06-01, Disponible en: http://www.omg.com/mda, 2003.
2. Object Management Group, Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification, OMG Document Number: formal/2011-01-01, Standard Document URL: http://www.omg.org/spec/QVT/1.1/PS/, últimoaccesoNoviembre 2012.
3. Object Management Group Meta Object Facility (MOF) Core Specfication OMG Available Specication. Versión 2.0. formal/06-01-01, http://www.omg.org/docs/formal/06-01-01.pdf, último acceso Febrero 2013.
4. Jacobson, I. El Proceso Unificado de Desarrollo de software. Addison-Wesley, EE.UU., 2000.

5. Booch G., Rumbaugh J., Jacobson I., The Unified Modeling Language. Addison Wesley, Second Edition. 2005.

6. Ariel Arsaute, Marcelo Uva, Fabio Zorzan, y otros, “Hacia una integración de MDA y el Proceso Unificado a través de reglas de transformación QVT”. 41 Jornadas Argentinas de Informática – JAIIO 2012.

7. Object Management Group, http://www.omg.org,últimoaccesoAbril 2013.

8. Object Management Group Object Constraint Language Version 2.0. OMG DocumentNumber: formal/06-05-01, Standard Document URL:http://www.omg.org/cgi-bin/doc?omg/03-06-01, último acceso Marzo 2013.

9. ikv++: medini QVT. http://www.ikv.de/, ultimo acceso Agosto 2012.

10. ORACLE, Java SE Documentation, Java RemoteMethodInvocation URL: http://docs.oracle.com/javase/6/docs/technotes/guides/rmi/index.html, último acceso Noviembre 2012.

11. Frank Buschmann&otros, Pattern-Oriented Software Architecture, Wiley; Volume 1 edition (August 8, 1996).

12. Erich Gamma, Richard Helm, Ralph Johnson, John Vlissedes. Design Patterns, Elements of Reusable Object/Oriented Software- Addison-Wesley, 1995.

13. Fowler, Martin, Patterns of Enterprise Application Architecture, Addison-Wesley, 2003.

Modeling Complex Mobile Web Applications from UI Components – Adding Different Roles and complex Database Design

Pablo Vera[1], Claudia Pons[2], Carina González González[3], Daniel Giulianelli[1], Rocío Rodríguez[1]

[1] National University of La Matanza
Department of Engeniering and Technological Research
San Justo, Buenos Aires, Argentina
{pvera, dgiulian, rrodriguez }@ing.unlam.edu.ar
[2] National University of La Plata
LIFIA – Research and Education Laboratory on Advance Computing
La Plata, Buenos Aires, Argentina
cpons@lifia.info.unlp.edu.ar
[3] La Laguna University
Department of Systems Engineering and Automation,
Architecture and Computer Technology
La Laguna, España
cjgonza@ull.es

Abstract. Component Based Hypermedia Design Method (CBHDM) is a modeling methodology that allows creating mobile web applications by designing and configuring user interface components. Starting from models this methodology performs two transformations to finally generate the application source code using the MDA approach. In order to configure the user interface components this methodology creates a custom language that´s powerful enough for designing complex applications. This paper shows how to configure components for allowing complex database design and also includes a new feature on the model supporting different user roles assigning different screens to operate the system.

Keywords: MDA, Mobile Web Applications, Mobile, UML, User Roles

1 Introduction

MDA (Model Driven Architecture) [1] is an approach for developing systems by building models and generating the application source code automatically or semi automatically by following some transformations steps. In order to be able to model complex systems the modeling methodology must support advanced capabilities like complex queries over the data model and assigning different views to different user roles.

Several methodologies use MDA approach to model web applications starting from the conceptual model and defining the navigational design. Some of them also include the desire capabilities to support advance modeling such as:

- Object Oriented Hypermedia Design Method (OOHDM) [2] allows querying data by using a sql like syntax over the objects. This syntax is used in the class definition to set related properties. For assigning different behavior for different roles a navigational model must be done for each role, defining a view of the conceptual model for each role.
- Web Modeling Language (WEBML) [3] includes an Object Query Syntax for accessing related data. It also includes support for users and groups allowing the definition of pages that will be visible by defining a site view for each group.
- Engineering Web Applications Using Roles [4] discuss role modeling in web engineering and proposed a notation for assigning roles to conceptual and navigation models of different methodologies.
- A MDA Approach for Navigational and User Perspectives [5] models roles with UML actors and hierarchy and then defines zones where those roles can operate. For each zone a navigation diagram is created.

In this paper we present a modeling methodology named "Component Based Hypermedia Design Method (CBHDM)" [6], with new features that allow creating mobile web applications through user interface components. Thus, the paper is organized as following: Section 2 will briefly introduce the methodology. Section 3 will explain the new features added to the methodology for supporting different user roles modeling. Section 4 will explain how to use the configuration capabilities of the methodology to support complex data design and query. Finally section 5 will show the conclusions and future work.

2 Component Based Hypermedia Design Method (CBHDM)

CBHDM is a design methodology for designing mobile web application. It´s based on a conservative extension of UML. It adds some necessary characteristics on class and component diagrams allowing a detailed modeling. Those models will have all necessary information for automatically deploying an application source code by using the MDA approach.

The methodology starts from the UML Class diagram where the system entities are defined using some stereotypes that will be used on the final transformation to facilitate the process of generating the database script, and for identifying and describing entities on the system. Later a transformation tool uses the class diagram to automatically generate a Component Diagram that the designer will modify and complete with the desire behavior of the interface. An additional UML State chart diagram could be used to define object states sequences that later will be checked on the system business logic.

The last step consists in using the transformation tool for a second time with all models as input. The result will be a database script and a fully functional application source code. Figure 1 shows the different stages of the methodology and also remarks

the steps that requires user participation. More details about the methodology can be found in [6].

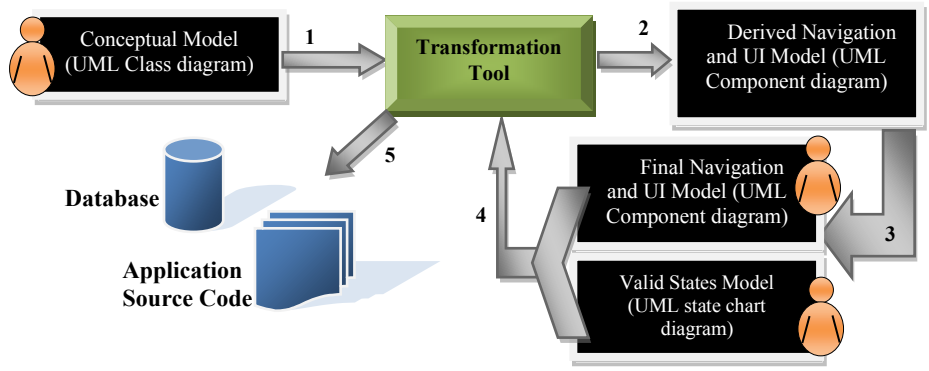


Fig. 1. CBHDM Methodology Stages

The power of CBHDM lies in the ability to configure pre-defined components.

The configuration is performed using tagged values with a detailed semantic for each value. The semantics is declared on a BNF that defines the configuration language.

This configuration language is powerful enough for customizing the components and also for accessing the data present on the conceptual modeling. It includes several functions that allow complex data resolution and query.

3 Roles

In order to allow defining different functionalities according to the role of the logged user a new parameter was added to the link function, the `RoleCondition`. The link function is used to show a link on the user interface to navigate to other component; it's the base of the navigation system and of the `MainMenu` component. A secondary function called `OptionalLink` was present allowing modeling links that are only visible if the condition is accomplished. In order to give more power to the system and to separate concerns the new role condition parameter was added instead of using the condition already present on the `OptionalLink` function giving more configuration power and encouraging clarity. The new parameter was added on both functions, `Link` and `OptionalLink`.

The role condition adds a rule that must be checked to determine the visibility of the link. This allows showing some links only to specific user roles.

The new form of the link function is:

```
<Link>::='Link('<LinkText>','<BrowsableComponent>','
<OptionalLinkParameters>','<OptionalAccessKey>','
<OptionalRoleCondition> ')'
```

And the new form of the OptionalLink function is:

```
<OptionalLink>::='OptionalLink''('<LinkText>','
<BrowsableComponent>','<OptionalLinkParameters>','
<OptionalAccessKey>','<LinkCondition>','
<OptionalRoleCondition>')'
```

The condition will be automatically related with the logged user, so the starting point must be the class representing the logged user. This approach adapts to different ways of representing the role assignment security that are next explained.

3.1 Direct Role on user class.

When role assignment is directly given by a property on the user class with a boolean property like in the example of the figure 2.

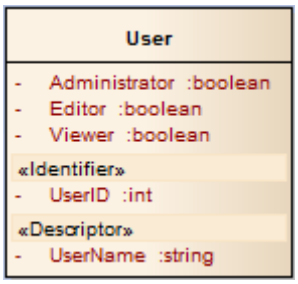


Fig. 2. Class with role assignment as property

A link that must be seen only by an administrator in the example of figure 2 will have the following Role Condition:

```
User.Administrator = true
```

The link can also be assigned to more than one role aggregating conditions, for example the following Role Condition will make the link visible either for administrators or editors users:

```
User.Administrator = true OR User.Editor = true
```

3.2 Unique Role with related class.

If the user has a unique Role the more common approach will be a property on the user class related to the role class. The Role class will establish the different roles on the system by enumeration values. So the designer could refer to those values to configure the access. An example of this approach can be seen in figure 3.

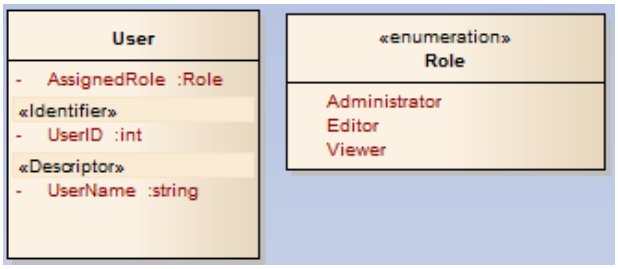


Fig. 3. Unique Role Assignment schema

A role condition restricting the visibility only to the administrator role will be:

```
User.Role = Role.Administrator
```

3.3 Several Roles with related class.

If the user can have more than one role and the roles are on a separate enum class. The conceptual model will have a class to assign security like the example in figure 4.

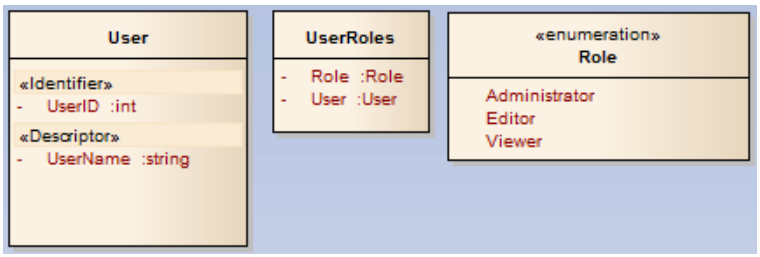


Fig. 4. Separate class for role assignment

In this case the main class of the condition is not the user class, so we must explicitly configure the field relating to the logged user like the code below:

```
UserRoles.User=LOGGEDUSER AND UserRoles.Role=Role.Editor
```

4 Complex Data Base Design

When modeling a system, a good database design is a key point for avoiding redundancy and for obtaining a correct system performance when accessing data.
CBHMD automatically generates the database from the conceptual model, so each class will be transformed on a database table. On each table a number of operations will be carried out to improve performance:

- The primary key will be set in the property marked as Identifier
- An index will be created for each property related to another table, for improving joins

– An unique index will be created for the descriptor property to avoid duplicated values

So the designer must create the class diagram thinking on the database design, knowing that later, data could be accessed due to a powerful configuration language for components. This will allow for example creating log records on a separate table when updating a table and accessing data by complex querying.

In order to illustrate the possibilities of the configuration language some examples will be shown next. All examples will be based on the conceptual model of the figure 5 related to a mobile system for registering trips in a taxi company. Note that CBHDM models class relationships by adding a property with the type of the related entity like in database design.

Fig. 5. Conceptual Model for registering trips on a taxi company

4.1 Complex Querying

CBHDM configuration language allows the use of object notation for accessing data on related classes. But also includes several functions to access indirect information present on the model. Those functions are Sum, Count, Exist, Not Exist, Eval and Retrieve.

Sum.

This function allows obtaining the sum of some property on a related table. For example getting the sum of the prices of trips finished of the logged driver:

```
Sum (Trip.Price, Trip.Driver = LOGGEDUSER AND
Trip.TripStatus = TripStatus.Finished)
```

The first parameter of the function is the property to be summed, the second are the conditions applied prior performing the operation.

Count.

It allows obtaining the quantity of objects that fulfills the condition. For example the following code gets the number of trips finished by the logged driver:

```
Count (Trip, Trip.Driver = LOGGEDUSER AND
Trip.TripStatus = TripStatus.Finished)
```

In this case the first parameter is the entity where the objects must be counted.

Exists and Not Exists.

These functions return a Boolean value to check if any object with a given condition exists. For example the following code creates a link to start a previously accepted trip only if a previous trip was not started before:

```
OptionalLink("Start", cpnStartTrip, "ObjectID =
TripID", TripStatus=TripStatus.Accepted AND not Exist
(TripStatus.Started));
```

The entity where to check the existence is not defined in this case because the link is part a component where the main entity was configured as Trip, so all related operations will be done on the Trip class. Otherwise the full syntax of the condition will be:

```
Trip.TripStatus=TripStatus.Accepted AND not Exist
(Trip.TripStatus.Started)
```

Eval.

It is a function to evaluate a condition and returning a Boolean value. For example in a table if we need to show inside a column that the trip price was 0 (bonus trip) we can use the eval function to check it:

```
Eval(Trip.Price = 0)
```

If the trip price was 0 it will show the “true” word in a column that for example could be labeled as “Bonus Trip”.

Retrieve.

When working with log classes is usually more efficient retrieving information from those classes instead of duplicating data on source class. So the retrieve function will be used to go and get the related data. For example if we want to show the initial request date of the trip in a grid we need to access the log table and find the initial status of the trip:

```
Retrieve (min(EventDateTime),
TripLog.TripStatus=TripStatus.Pending);
```

The first parameter is the property to retrieve and the second the condition. This function must be used in the context of a list where the id of the trip is taken from the row being displayed and this is an implicit filter for the TripLog.Trip field.

4.2 Partially Updating records

In several systems an object class goes through different states, and on each state not all its properties must be able to be updated. For that reason CBHDM adds the `UpdateView` component that was specially created for allowing that partial update

of the object. For example, the driver has a list with available trips and he wants to confirm that he will perform that trip. In that case all trip information is already set and the driver will only change the trip status and eventually fill a text with some remarks. The `UpdateView` component allows showing some properties in only read mode and other in edition mode. In the example of table 1 the remarks are configured to be editable only by the user and the status is automatically changed when updating the object.

Table 1: Tagged values for "Accept Trips“ component of type UpdateView

Tag	Value
Id	cpnAcceptTrip
Navigation	Link(“Main Menu”, cpnMainMenu,,0); Link("Back", cpnPendingTrips,,9);
MainEntity	Trip
DisplayProperties	CustomerAddress; DestinationAddress; Retrieve(min(EventDateTime),TripLog, TripStatus=TripStatus.Pending,"Request Date"); CustomerPhone
UpdateProperties	TripLog.Remarks
DefaultValuesUpdate	TripStatus = TripStatus.Accepted; Driver = LOGGEDUSER;

4.3 Creating additional records

Usually when an object is created or updated a new record in a related table must be created. This approach is very common for a log class that records the changes made on some particular class. CBHDM includes a special tagged value called `CreateEntity` that creates an object on the entity configured on the value of this tag. For example, if the system must keep record of the change of the status of the trip that was configured on table 1, the `CreateEntity` tag can be used. Table 2 shows the complete component configuration adding a configuring log record creation. The values of the newly created object are configured with object notation as can be seen in the value of the `DefaultValuesUpdate` tag of table 2.

Table 2: Tagged values for "Accept Trips“ component of type UpdateView with log record

Tag	Value
Id	cpnAcceptTrip
Navigation	Link(“Main Menu”, cpnMainMenu,,0); Link("Back", cpnPendingTrips,,9);
MainEntity	Trip
CreateEntity	TripLog
DisplayProperties	CustomerAddress; DestinationAddress; Retrieve(min(EventDateTime),TripLog, TripStatus=TripStatus.Pending,"Request Date"); CustomerPhone
UpdateProperties	TripLog.Remarks
DefaultValuesUpdate	TripStatus = TripStatus.Accepted; Driver = LOGGEDUSER;

	`TripLog.Driver = LOGGEDUSER;` `TripLog.EventDateTime = NOW`

For the same purpose two tagged values were added to the CRUD component that allows creating and updating class objects: `CreateEntityOnCreate` and `CreateEntityOnUpdate`. This allows adding related records when creating an object or when updating if `UpdateView` component is not used.

5 Conclusions and Future Work

CBHDM allows modeling systems with all necessary information for automatic code generation. The use of functions to retrieve data is powerful enough for accessing related records and for complex querying as shown in the examples above.

The new role property on links allows the designer to assign tasks to each role for increasing system security and to assign responsibilities without the need of creating a separate model for each role like the methodologies explained in section 1.

Finally the ability to create related records when updating or creating a main class is an essential characteristic that completes the model and allows for example keeping track of modification in a separate log table.

CBHDM methodology and language configuration is now completely defined and the transformation tool is being developed. Future work consists on finishing the mentioned development and validating the results by performing modeling of different complexity.

6 References

1. Kleppe A., Warmer J., Bast W. “MDA explained: the model driven architecture: practice and promise”. Addison-Wesley Professional (2003)
2. Schwabe D. y Rossi G. “An object oriented approach to Web-based applications design”. Theor. Pract. Object Syst. Volume 4, Issue 4 (1998), pp 207-225.
3. Ceri S., Fraternali P., Bongio. “Web Modeling Language (WebML): a modeling language for designing Web sites”, Computer Networks, Volume 33, Issues 1–6, (2000), pp 137-157.
4. Rossi G. Nanard J., Nanard M and Koch Nora, “Engineering Web Applications Using Roles“, Journal of Web Engineering, Vol. 6, No.1 (2006
5. Gonzales M, Casariego J., Bareir J., Cernuzzi L, Pastor O. “A MDA Approach for Navigational and User Perspectives”, Special issue of best papers presented at CLEI 2010 (2011)
6. Vera P., Pons C. Gonzales C, Giulianelli D., Rodriguez R. “MDA based Hypermedia Modeling Methodology using reusable components”, XVIII Congreso Argentino de Ciencias de la Computación (2012)

Un Análisis Experimental de Tipo de Aplicaciones para Dispositivos Móviles

Lisandro Delía[1], Nicolás Galdamez[1], Pablo Thomas[1], Patricia Pesado[1]

1 Instituto de Investigación en Informática LIDI. Facultad de Informática.
Universidad Nacional de La Plata. Argentina
{ldelia, ngaldamez, pthomas, ppesado}@lidi.info.unlp.edu.ar

Resumen. El auge de los dispositivos móviles ha generado nuevos desafíos para los ingenieros de software. Las capacidades técnicas ofrecidas, así como sus restricciones, plantean un escenario fértil, pero complejo. Existen diferentes alternativas de desarrollo de una misma aplicación para un dispositivo móvil. En este trabajo se presentan los enfoques de desarrollo de software existentes, sus características más destacadas, y un caso experimental que permite analizar las ventajas y dificultades de cada enfoque.

Palabras claves: dispositivos móviles, aplicaciones móviles nativas, aplicaciones móviles híbridas, aplicaciones móviles web.

1 Introducción

Los dispositivos móviles forman parte de la vida cotidiana y son cada vez más sofisticados, su poder de cómputo genera posibilidades hasta hace años no pensadas.

La creciente demanda de software específico para estos dispositivos ha generado nuevos desafíos para los desarrolladores, ya que este tipo de aplicaciones tiene sus características propias, restricciones y necesidades únicas, lo que difiere del desarrollo de software tradicional.

La computación móvil se puede definir como un entorno de cómputo con movilidad física. El usuario de un entorno de computación móvil será capaz de acceder a datos, información u otros objetos lógicos desde cualquier dispositivo en cualquier red mientras está en movimiento [1].

Las particularidades de este entorno incluyen: alto nivel de competitividad, tiempo de entrega necesariamente corto y la dificultad adicional de identificar los stakeholders y sus requerimientos.

Las aplicaciones se generan en un entorno dinámico e incierto. Generalmente, son pequeñas, no críticas, aunque no menos importantes. Están destinadas a un gran número de usuarios finales y son liberadas en versiones rápidas para poder satisfacer las demandas del mercado [2].

El desarrollo de aplicaciones móviles es, actualmente, un gran desafío, dado las demandas específicas y las restricciones técnicas de un entorno móvil [3], tales como dispositivos con capacidades limitadas, pero en evolución continua; varios estándares, protocolos y tecnologías de red, necesidad de operar sobre diferentes plataformas,

requerimientos específicos de los usuarios y las exigencias estrictas en tiempo del mercado.

Estos dispositivos tienen características físicas distintivas, entre las cuales se destacan su tamaño, peso, tamaño de pantalla, su mecanismo de ingreso de datos y su capacidad de expansión. Además, los aspectos técnicos, incluyendo el poder de procesamiento, espacio de memoria, autonomía de batería, sistema operativo, entre otros, tienen un rol esencial. Todas estas características deben ser cuidadosamente consideradas en el desarrollo de aplicaciones [4].

En la corta historia del desarrollo de software las plataformas de hardware y software han evolucionado en forma constante, pero nunca antes fue tan masivo el poder de cómputo que tienen las personas en sus manos, puntualmente a través del uso de dispositivos móviles. Esto conduce a nuevos desafíos y junto a ellos al crecimiento de la Ingeniería de Software como disciplina, acompañando esta evolución.

En este trabajo se presenta un estudio comparativo de tipos de aplicaciones para dispositivos móviles, a partir de una caso experimental desarrollado para la plataforma educativa WebUNLP [6]. En la sección 2 se detallan las características más salientes de los diferentes tipos de aplicaciones para dispositivos móviles. Posteriormente se presenta el proceso de desarrollo de los diferentes tipos de aplicaciones para el caso experimental utilizado. Finalmente, se expresan conclusiones y trabajos futuros.

2 Tipo de aplicaciones para dispositivos móviles

En los últimos años el mercado de los dispositivos móviles, en especial smartphones, ha mostrado un crecimiento notable tanto en Argentina como en todo el mundo. En particular, en nuestro país, las plataformas que más han crecido son Android e iOS [8] [9].

Cada una de estas plataformas cuenta con una infraestructura de desarrollo particular.

El principal reto para los proveedores de aplicaciones es proporcionar soluciones para todas las plataformas, pero tiene un alto costo [11].

La solución ideal a este problema es crear y mantener una única aplicación para todas las plataformas. El desarrollo multiplataforma tiene como objetivo mantener la misma base de código para diversas plataformas. De esta forma el esfuerzo y costo de desarrollo se reduce notablemente.

A continuación se presentan tres enfoques para desarrollo de aplicaciones para dispositivos móviles: un enfoque nativo y dos enfoques multiplataforma (web e híbrido).

2.1 Aplicaciones Web

Las aplicaciones web para móviles son diseñadas para ser ejecutadas en el navegador del dispositivo móvil. Estas aplicaciones son desarrolladas utilizando HTML, CSS y JavaScript, es decir, la misma tecnología que la utilizada para crear sitios web.

Una de las ventajas de este enfoque es que los dispositivos no necesitan la instalación de ningún componente en particular, ni la aprobación de algún fabricante para que las aplicaciones sean publicadas y utilizadas. Solo se requiere acceso a internet. Además, las actualizaciones de la aplicación son visualizadas directamente en el dispositivo, ya que los cambios son aplicados sobre el servidor y están disponibles de inmediato. En resumen, es rápido y fácil de poner en marcha.

La principal ventaja de este tipo de aplicación es su independencia de la plataforma. No necesita adecuarse a ningún entorno operativo. Solo es necesario un navegador.

Por contrapartida, esto disminuye la velocidad de ejecución y podrían llegar a ser menos atractivas que las aplicaciones nativas. Además, podrían tener baja performance por problemas de conectividad. Finalmente este tipo de aplicaciones no pueden utilizar todos los elementos de hardware del dispositivo, como por ejemplo, cámara, GPS, entre otros.

2.2 Aplicaciones Nativas

Las aplicaciones nativas son aquellas que se conciben para ejecutarse en una plataforma específica, es decir, se debe considerar el tipo de dispositivo, el sistema operativo a utilizar y su versión.

El código fuente se compila para obtener código ejecutable, proceso similar que el utilizado para las tradicionales aplicaciones de escritorio.

Cuando la aplicación está lista para ser distribuida debe ser transferida a las App stores (tiendas de aplicaciones) específicas de cada sistema operativo. Estas tienen un proceso de auditoría para evaluar si la aplicación se adecúa a los requerimientos de la plataforma a operar. Cumplido este paso, la aplicación se pone a disposición de los usuarios.

La principal ventaja de este tipo de aplicaciones es la posibilidad de interactuar con todas las capacidades del dispositivo (cámara, GPS, acelerómetro, agenda, entre otras). Además no es estrictamente necesario poseer acceso a internet. Su ejecución es rápida, puede ejecutarse en modo background y notificar al usuario cuando ocurra un evento que necesite su atención.

Claramente estas ventajas se pagan con un mayor costo de desarrollo, pues se debe utilizar un lenguaje de programación diferente según la plataforma. Por ende, si se desea cubrir varias plataformas, se deberá generar una aplicación para cada una de ellas. Esto conlleva a mayores costos de actualización y distribución de nuevas versiones.

2.3 Aplicaciones Híbridas

Las aplicaciones híbridas combinan lo mejor de los dos tipos de aplicaciones anteriores. Se utilizan tecnologías multiplataforma como HTML, Javascript y CSS, pero se puede acceder a buena parte de las capacidades específicas de los dispositivos.

En resumen, son desarrolladas utilizando tecnología web y son ejecutadas dentro de un contenedor web sobre el dispositivo móvil.

Entre las principales ventajas de esta metodología se pueden mencionar la posibilidad de distribución de la aplicación a través de las tiendas de aplicaciones, la reutilización de código para múltiples plataformas y la posibilidad de utilizar las características de hardware del dispositivo.

Una de las desventajas es que, al utilizar la misma interfaz para todas las plataformas, la apariencia de la aplicación no será como la de una aplicación nativa. Finalmente la ejecución será más lenta que la ejecución en una aplicación nativa.

3 Caso experimental: WebUNLP

3.1 Descripción del problema

WebUNLP es un entorno virtual de enseñanza y aprendizaje que permite a los docentes mediar sus propuestas educativas. Alumnos y docentes pueden encontrarse en ese espacio para compartir materiales de estudio, comunicarse y generar una experiencia educativa en forma virtual [6].

Actualmente, WebUNLP cuenta con una versión web enfocada a computadoras de escritorio o portátiles, pero no está adaptada para ser utilizada desde dispositivos móviles.

El desarrollo planteado en este trabajo consiste en extender WebUNLP, con la construcción de una aplicación móvil que permita acceder a determinadas funcionalidades del sistema a través de un dispositivo móvil. El enfoque propuesto incluye un análisis de la misma solución, comparando el desarrollo nativo, web e híbrido, a fin de establecer cuál de ellos es conveniente.

Como ocurre con cualquier desarrollo de software, la construcción de una aplicación móvil implica tener claramente definido su propósito y cuáles requerimientos debe cumplir. En particular, para el caso de software para dispositivos móviles resulta esencial tener objetivos más específicos que en su versión de escritorio [7].

Para el caso específico de WebUNLP se realizó un desarrollo incremental de una aplicación móvil, y esta primera versión se enfocó en una de sus herramientas de comunicación: la cartelera. Esta herramienta permite comunicar las novedades de un curso, como, por ejemplo, el cambio de horario de una cursada, recordar las fechas de entrega de un trabajo práctico, entre otras [6].

3.2 Análisis

Uno de los primeros interrogantes planteados fue la elección de la plataforma. En términos de mercado los sistemas operativos que predominan en Argentina son Android e iOS [8][9], con lo cual se decidió dar soporte a estos dos sistemas operativos.

Se analizaron los requerimientos funcionales y no funcionales de forma aislada a la plataforma y luego de forma específica para cada una de ellas.

A continuación, se presentan algunos requerimientos a cumplir por la aplicación:

- El usuario debe poder ingresar a la aplicación con las mismas credenciales que las utilizadas para acceder a la versión web.
- El usuario debe poder acceder a la cartelera de todos los cursos en los que participa, ya sea como docente o alumno.
- El usuario debe recibir una notificación en su dispositivo cuando una novedad es publicada en la cartelera. Este requerimiento no se puede cumplir en la versión web accesible desde computadoras de escritorio y/o portables.
- El usuario debe tener la misma experiencia de uso en todas las plataformas operativas
- La aplicación web existente debe estar sincronizada con la aplicación móvil a desarrollar, esto significa que cualquier cambio realizado desde la aplicación móvil debe verse reflejado en la versión web y viceversa.

3.3 Diseño

Para satisfacer los requerimientos planteados en el punto anterior, a excepción de la notificación, el diseño de la aplicación móvil web consiste en una réplica de lo ofrecido por WEBUNLP, adaptándose solamente la interfaz al tamaño de pantalla del dispositivo móvil.

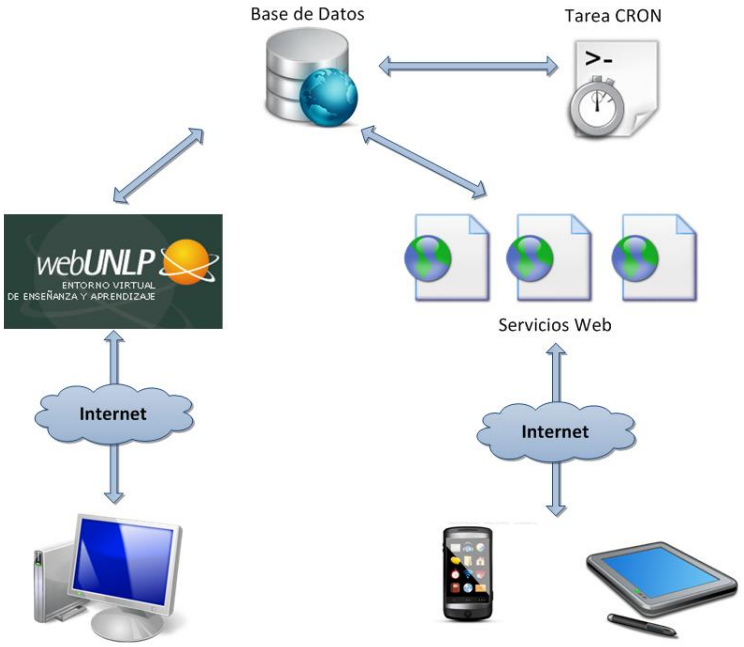


Figura 1 - Arquitectura genérica para aplicación nativa e híbrida

Para el diseño de una aplicación móvil nativa y una híbrida, la situación es más compleja. En la figura 1 se presenta la arquitectura genérica de todos los componentes que participan en este escenario de desarrollo. En ella se observa el acceso desde una PC a WebUNLP, y el acceso desde dispositivos móviles (celular y tablet) a la información de WebUNLP, mediante servicios web. Para el desarrollo de los servicios web se diseñó una API (Application Programming Interface) Restful dada su simpleza, escalabilidad e interoperabilidad [14] [15].

Asimismo existe una tarea programada (Cron) de ejecución intermitente a intervalos regulares de tiempo, la cual notifica a los dispositivos móviles correspondientes cuando se crea una novedad en la cartelera de WebUNLP. El Cron identifica el sistema operativo del dispositivo a notificar y genera dicha notificación.

En cuanto a la interfaz gráfica, se planteó un diseño independiente de la plataforma para analizar los aspectos de usabilidad de la aplicación, y luego se realizaron los ajustes necesarios para cada tipo de aplicación.

En las interfaces diseñadas la navegación es en serie, en el orden en que se presentan en el mockup [13] de la figura 2.

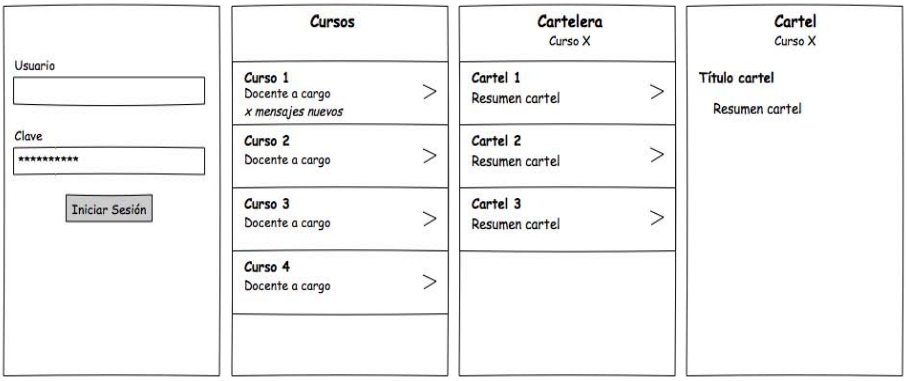


Figura 2 - Mockup independiente del tipo de aplicación

3.4 Desarrollo

3.4.1 Aplicación Nativa para Android

El desarrollo de aplicaciones para la plataforma Android requiere disponer de un JDK (Java Development Kit) y su entorno de programación, conocido como Android SDK (Software Development Kit). Este último provee librerías y herramientas necesarias para construir, testear y depurar aplicaciones para Android.

El desarrollo de la aplicación de WebUNLP para Android se realizó según las convenciones adoptadas por su comunidad, porque aplica buenas prácticas en la construcción de las interfaces y la lógica detrás de ellas, el acceso a datos y a servicios web. Para cumplir el requerimiento de las notificaciones en el dispositivo correspondiente, el Cron genera la notificación mediante el uso del servicio GCM

(Google Cloud Messaging) [16]. En la figura 3, se presentan las interfaces de la aplicación desarrollada.

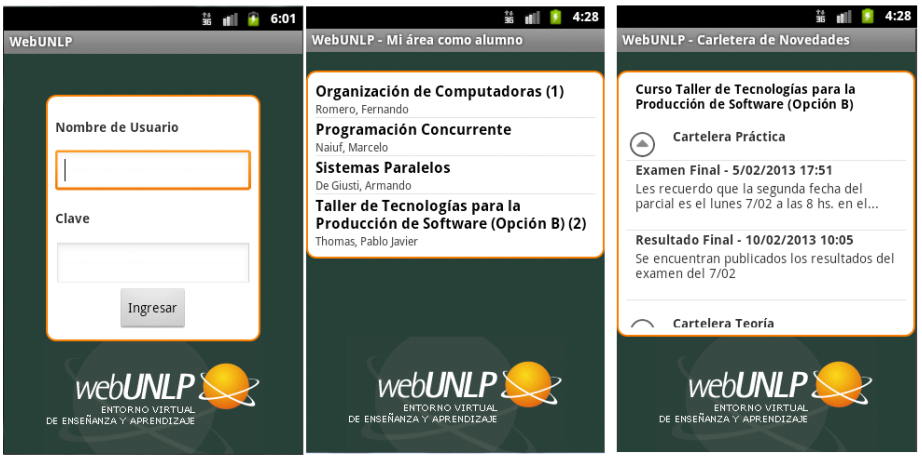


Figura 3 - Aplicación nativa para Android

3.4.2. Aplicación Nativa para iOS

La plataforma Apple iOS está basada en un modelo propietario, es por ello que el desarrollo de una aplicación nativa iOS implica contar con una Apple Mac corriendo OS X con Xcode instalado. El lenguaje de programación principal es Objective C. Xcode es el entorno de desarrollo de Apple para todos sus dispositivos y es el encargado de proporcionar el iOS SDK con las herramientas, compiladores y frameworks necesarios. Además, Xcode viene integrado con simuladores para dispositivos iOS (iPhones y iPads) que facilitan las etapas de prueba del sistema desarrollado.

Para los aspectos referentes a la interfaz gráfica y la interacción con el usuario, se siguieron las convenciones propuestas por Apple [10] para lograr una mejor integración de la aplicación con el sistema operativo y mejorar la experiencia del usuario.

Por último, para satisfacer el requerimiento de notificaciones, el Cron notifica al dispositivo iOS correspondiente mediante el uso del servicio APNs (Apple Push Notification Service) [17]. En la figura 4, se presentan las interfaces de la aplicación desarrollada.

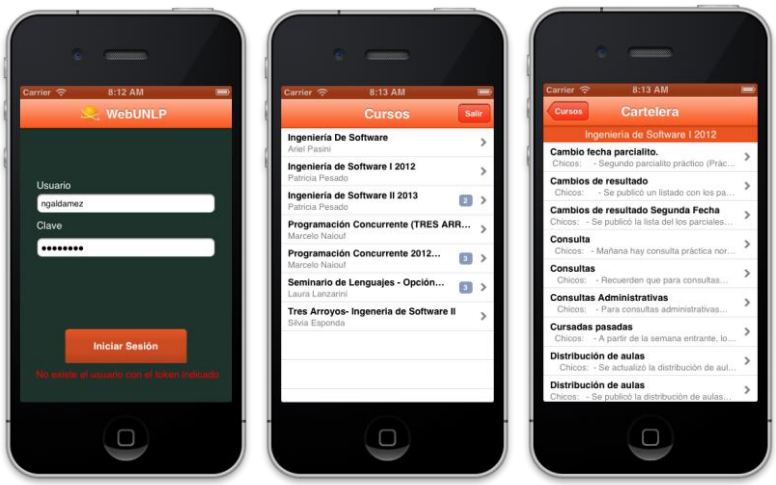


Figura 4 - Aplicación nativa para iOS

3.4.3 Aplicación Híbrida

Para la construcción de la aplicación WebUNLP híbrida se utilizó el framework PhoneGap [19], el cual permite desarrollar aplicaciones móviles que utiliza tecnologías comunes a todos los dispositivos: HTML5, CSS y Javascript.

Asimismo, se utilizó el framework Javascript denominado Jquery Mobile [20] para lograr interfaces con apariencia y comportamiento consistente a través de las diferentes plataformas móviles.

Con el objetivo de implementar el patrón de diseño MVC (Modelo, Vista, Controlador) se utilizó la librería Backbone.js [21].

Por último, para satisfacer el requerimiento de notificaciones, se utilizó el plugin Pushwoosh [18].

En la figura 5, se presentan las interfaces de la aplicación desarrollada.

3.4.4 Aplicación Web

Finalmente, se desarrolló una aplicación web capaz de acceder a la cartelera de WebUNLP. La misma se encuentra disponible para cualquier dispositivo móvil que cuente con un navegador que soporte las características utilizadas para el desarrollo: HTML5, CSS3, Javascript.

Como la velocidad de transmisión/recepción de datos de un dispositivo móvil a través de WiFi, y en particular 3G, es inferior a la velocidad de una computadora de escritorio, la versión web de la cartelera de WebUNLP es liviana y gran parte de los requerimientos son implementados a través de Ajax [12] para evitar, ante algún cambio, la recarga de la página completa.

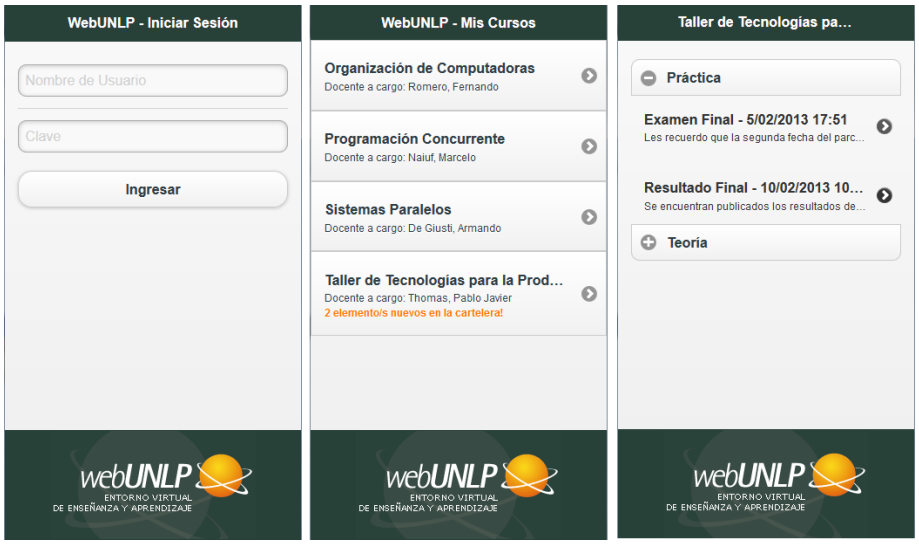


Figura 5 - Aplicación híbrida

4 Conclusiones

Inicialmente los dispositivos móviles fueron pensados y diseñados con un propósito especial. A través de los años, el crecimiento tecnológico ha permitido incorporar funcionalidades adicionales, lo que ha posibilitado expandir el marco de uso.

Actualmente, el poder de cómputo subyacente en una gran variedad de dispositivos móviles ha generado nuevas posibilidades, lo que constituye un reto para los ingenieros de software.

Es difícil establecer afirmaciones rotundas en este entorno de evolución vertiginoso y continuo. Es claro que, por el momento, existen tres posibilidades de desarrollo de una misma aplicación, y la incursión en cada una de ellas ha permitido establecer algunas conclusiones.

Seleccionar un dominio para desarrollar un caso experimental tiene sus particularidades. La novedad de disponer una aplicación móvil no es motivo suficiente que justifique tal desarrollo. Es necesario satisfacer un conjunto de requerimientos claramente planteados.

WebUNLP es un entorno virtual de enseñanza y aprendizaje utilizado por diversos cursos de grado y postgrado de la Universidad Nacional de La Plata. Por ende, la cantidad de usuarios beneficiados por acceder desde cualquier lugar a este entorno es importante.

De esta manera, se optó por elegir este espacio virtual para ser replicado en una versión móvil que no sólo permita acercarlo a sus usuarios, sino que además amplíe sus funciones, en este caso puntualmente para la cartelera.

Actualmente, se han desarrollado los tres tipos de aplicación móvil (web, nativa e híbrida) con el mismo propósito y para satisfacer el mismo conjunto de requerimientos.

Se destaca la gran simpleza de generar la versión web, dado que se utilizan las mismas herramientas tecnológicas que las usadas para el desarrollo de cualquier aplicación web tradicional. La principal diferencia en este sentido radica en la limitación de espacio en la pantalla del dispositivo. No obstante, el mayor contraste está dado en que no es posible acceder a todas las capacidades de hardware del dispositivo, lo que impidió implementar uno de los requerimientos, probablemente, el más interesante: la notificación de novedades de la cartelera de WebUNLP al usuario.

Por otra parte, las versiones nativas han cumplido todos los requerimientos. Aunque la mayor desventaja consiste en la no portabilidad, lo que implicó un desarrollo específico para la plataforma que se desee cubrir. En este trabajo se han presentado los desarrollos para Android e iOS, sistemas operativos más usados en Argentina, con un costo inherente mayor.

Finalmente, la versión híbrida ha logrado conjugar la simpleza del desarrollo web con el uso de todas las capacidades del dispositivo. Este tipo de enfoque pretende suplir las desventajas de los dos enfoques previos, hecho que lo posiciona como la elección prima facie, siempre condicionada por los requerimientos específicos a cumplir.

Como prueba de uso, recientemente se puso disponible la aplicación en su modo nativo, que logró captar la atención de los usuarios, traducida en el pedido de incorporación de nuevas funciones presentes en WebUNLP.

5 Trabajo futuro

Se pretende expandir a corto plazo algunos requerimientos cumplidos por WebUNLP que no se satisfacen en su versión móvil, como, por ejemplo, la mensajería y el foro, entre otros. Además, se espera incorporar nuevos requerimientos tales como un servicio de chat.

Desde el punto de vista del desarrollo, se analizarán alternativas para generar aplicaciones móviles híbridas mediante otros frameworks [22] [23].

Referencias

1. Talukder, A.K., Ahmed, H. and Yavagal,R.: *Mobile Computing, Technology, Applications, and Service Creation*. Second Edition. Tata McGraw-Hill. 2010.
2. Abrahamsson, P. *Mobile software development - the business opportunity of today*. Proceedings of the International Conference on Software Development, (pp. 20-23). 2005. Reykjavik.
3. Hayes, I. S. *Just Enough Wireless Computing*. Prentice Hall Professional Technical Reference . 2002. ISBN:0130994618
4. Abrahamsson P. et. al. , *Mobile-D: An Agile Approach for Mobile Application Development.* OOPSLA'04, Oct. 24–28, 2004, Vancouver, British Columbia, Canada.
5. Tracy, K.W., *Mobile Application Development Experiences on Apple's iOS and Android OS*, Potentials, IEEE, 2012

6. Sitio de WebUNLP. http://webunlp.unlp.edu.ar
7. Salmre, I.. *Writing Mobile Code Essential Software Engineering for Building Mobile Applications*. Addison Wesley Professional, 2005
8. http://gs.statcounter.com
9. http://www.mapbox.com/labs/twitter-gnip/brands/#4/-40.43/-63.62
10. *iOS Human Interface Guidelines*, http://developer.apple.com/library/ios/#DOCUMENTATION/UserExperience/Conceptual/MobileHIG/Introduction/Introduction.html
11. Raj R.,Tolety S.B. *A study on approaches to build cross-platform mobile applications and criteria to select appropriate approach*. India Conference (INDICON), 2012 Annual IEEE
12. https://developer.mozilla.org/en/docs/AJAX
13. http://es.wikipedia.org/wiki/Mockup
14. Richardson L. ,Ruby S., RESTful Web Services, O'Reilly Media, 2007.
15. http://msdn.microsoft.com/es-es/magazine/dd315413.aspx#id0070023
16. http://developer.android.com/google/gcm/index.html
17. http://developer.apple.com/library/mac/#documentation/NetworkingInternet/Conceptual/RemoteNotificationsPG/Chapters/ApplePushService.html#//apple_ref/doc/uid/TP40008194-CH100-SW9
18. http://devgirl.org/2012/12/04/easy-phonegap-push-notifications-with-pushwoosh/
19. http://phonegap.com/
20. http://jquerymobile.com/
21. http://backbonejs.org/
22. Digital Possibilities. Mobile Development Frameworks Overview http://digital-possibilities.com/mobile-development-frameworks-overview/
23. Markus Falk. Mobile Frameworks Comparison Chart, http://www.markus-falk.com/mobile-frameworks-comparison-chart/

Evolución Semántica de Glosarios en los Procesos de Requisitos

Gladys Kaplan[1], Jorge Doorn[1,2], Nora Gigante[1]

[1]Departamento de Ingeniería e Investigaciones Tecnológicas, UNLaM
[2]INTIA, Departamento de Computación y Sistemas, Facultad de Ciencias Exactas, UNCPBA
gladyskaplan@gmail.com, jdoorn@exa.unicen.edu.ar, noragigante@gmail.com

Resumen. El uso de glosarios en los procesos de requisitos está ampliamente difundido. Los glosarios construidos muy tempranamente en el proceso de requisitos y que son utilizados luego en todo el proceso, tienden a desactualizarse a medida que el proyecto evoluciona. Se ha comprobado que la mera planificación de un nuevo sistema de software, requiere crear documentos parar describir situaciones inexistentes o planificadas, las cuales no pueden ser descriptas con los términos registrados hasta el momento en el glosario. Estos términos nuevos o resignificados deben ser agregados al glosario. De lo contrario, un glosario construido con el propósito de reducir la ambigüedad se puede convertir, paradójicamente, en un factor que contribuye a incrementarla. En este artículo se analizan los cambios semánticos del vocabulario producidos durante el proceso de requisitos y se propone un mecanismo para elicitarlos y modelarlos.

Palabras Clave*:* Procesos de requisitos, evolución semántica, LEL de requisitos.

1 Introducción

La construcción de glosarios como parte del proceso de desarrollo del software en general y de la Ingeniería de Requisitos en particular [1][2][3][4][5] ha sido reconocida como una actividad necesaria para asegurar la comunicación y la comprensión de todo el conocimiento del proceso de requisitos. Estos glosarios son utilizados durante el proceso de requisitos sirviendo de referencia a toda la documentación generada. El uso de los glosarios facilita la elicitación de conocimiento, mejora la comunicación con los clientes-usuarios y reduce la ambigüedad de los documentos construidos. Sin embargo se ha prestado poca atención a la obsolescencia que pueden padecer estos glosarios ya que son construidos muy tempranamente en el proceso.

Es frecuente encontrar ambigüedad, conflictos, inconsistencias en el vocabulario los cuales tienen su origen en el propio Universo del Discurso (UdeD)[1]. La existencia de conflictos en el vocabulario del UdeD es habitualmente una realidad preexistente al proceso de desarrollo del software. Es así que la actividad de construcción de un glosario debe lidiar con estos conflictos, las que usualmente se manifiestan como homónimos y como sinónimos.

La construcción de documentos que describen los cambios del proceso del negocio como consecuencia de la inserción del nuevo sistema de software, son el factor determinante de la aparición de otros conflictos pre-existentes y con la particularidad de ser más insidiosos ya que son poco notados por los ingenieros de requisitos y poco notados por los clientes y usuarios que tienen acceso a esos documentos recién en la fase final de su construcción.

[1] UdeD: es todo el contexto en el cual el software se desarrolla e incluye todas las fuentes de información. Es la realidad acotada por el conjunto de objetivos establecidos por quienes demandan una solución de software"[6]

Es importante marcar que el hito en el cual se libera el glosario como un documento disponible, es lo que determina su vigencia. El glosario creado inicialmente es utilizado por el ingeniero de requisitos tanto como sea posible para describir las situaciones observables en el UdeD y también en las situaciones que ocurrirán cuando se instale el nuevo sistema de software. Pero en este último caso ese glosario no alcanza ya que se deben introducir nuevos términos para describir nuevas situaciones. Por lo tanto los documentos más avanzados del proceso, requerirán algunas definiciones diferentes a las registradas en el glosario inicial o mencionarán términos relevantes para la solución que directamente no están incluidos en dicho glosario. Es paradójico que el glosario que fue creado para reducir la ambigüedad de los modelos construidos se convierta, él mismo, en una fuente de ambigüedad. Es necesario que estos glosarios construidos tempranamente sean actualizados a lo largo del proceso de requisitos, de tal manera que conserven las consistencias con los documentos a los cuales le dan apoyo semántico.

Se debe remarcar que el UdeD tampoco preserva en forma inmutable el vocabulario registrado al inicio del proceso de requisitos. Los clientes o usuarios también modifican su vocabulario como consecuencia de participar en la planificación del contexto en el cual se desenvolverá el futuro sistema de software. A estos cambios en el vocabulario lo denominamos aquí, evolución semántica. Existe una controversia en la bibliografía, acerca de la utilización de la palabra evolución. Para algunos autores el concepto está ceñido a las transformaciones que sufren los modelos durante todo el proceso de construcción del software, desde el análisis, pasando al diseño y así sucesivamente. Este es el caso de [7] donde se denomina evolución a las transformaciones de los escenarios y otros documentos durante todo el ciclo de vida del desarrollo de software. Otros, se refieren a la evolución como los cambios que sufre el dominio producto de planificar la inserción de un nuevo sistema de software. El presente artículo adhiere a la última definición pero con una salvedad, los cambios del contexto sólo existen en los modelos del proceso de requisitos, ya que aún no han sido implementados en los procesos del negocio. A esta evolución se la denomina en el presente artículo **evolución *aparente***. La evolución semántica o real sucederá recién cuando el sistema de software se encuentre en ejecución.

El fenómeno de la perdida de vigencia de los glosarios, tema central de este trabajo, ha sido estudiado particularmente en el Proceso de Requisitos Basado en Escenarios [8]. La primera actividad de este proceso es generar un glosario denominado Léxico Extendido del Lenguaje (LEL) [9] [13]. Este glosario contiene la vista léxica del UdeD. En su primera versión, el LEL describe el vocabulario de los clientes-usuarios. Este documento resulta notoriamente apropiado para describir el proceso de negocio actual u observable. Es así que el LEL cumple un rol predominante para asegurar el proceso, evidenciando la siguiente sobrecarga de objetivos:

- describir las palabras o frases que son relevantes en el contexto o tienen un significado distintivo para los clientes o usuarios.
- reducir la ambigüedad en los artefactos generados durante todo el proceso de requisitos.

Estos objetivos se satisfacen en su totalidad en la primera etapa del proceso, donde se elicita conocimiento para comprender los procesos del negocio. A medida que se avanza en el proceso de requisitos el UdeD evoluciona, pero el LEL sigue reflejando el vocabulario del UdeD actual. Es así que los objetivos iniciales del LEL se ven debilitados produciendo una pérdida de consistencia entre los documentos generados para describir la solución, llegando al vínculo ERS-LEL. Si bien queda claro que el LEL debe evolucionar, no hay ninguna certeza de que estos cambios del vocabulario sean coherentes, y que finalmente

algunos términos nuevos o modificaciones sean coincidentes con la realidad. Lo que determina que puede existir un LEL para el proceso de requisitos y otro con el vocabulario real.

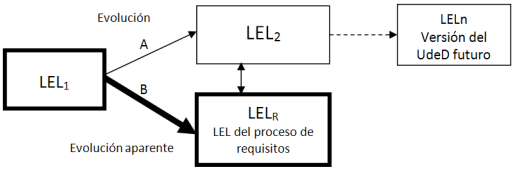


Figura 1 – Evolución del LEL

La Figura 1 muestra en el camino A, que existe un LEL relacionado con la evolución del UdeD. Este glosario registra las modificaciones del vocabulario muy lentamente pero genera un documento con una vigencia mucho más prolongada. En caso de crearse, no ofrece ninguna utilidad al proceso de requisitos ya que su generación excede al proceso de construcción del software. Por su parte el camino B, **LEL_R** o LEL de Requisitos, es aquel relacionado con el proceso de requisitos. Es significativamente diferente del LEL inicial e incorpora todos los términos necesarios para describir los escenarios futuros [14] y la Especificación de Requisitos de Software (ERS). Este documento debe construirse para mejorar la calidad de los documentos de requisitos, por tal motivo se lo denomina a esta evolución *aparente*, pero su vida es limitada ya que su utilidad decrece a lo largo del proceso de desarrollo. Sin embargo siempre es posible que deba ser consultado en etapas tardías del mismo. El LEL evolucionado y el LEL_R, son permeables y algunos términos se trasladan de uno a otro. Idealmente un gran involucramiento de los clientes y usuarios podría lograr que las diferencias entre ambos sean menores.
En la siguiente sección se describe la evolución del dominio. La sección 3 describe cómo la evolución conceptual del contexto impacta en el vocabulario utilizado para describir los escenarios futuros y los requisitos del software. En la sección 4 se realiza un análisis sobre las estrategias para construir el LEL_R. En la sección 5 se presenta la heurística de construcción del LEL_R. Y finalmente en la sección 6 se mencionan algunas conclusiones.

2 Evolución del Dominio

Los dominios cambian desde el mismo momento en el que se comienza a "pensar" en incorporar un nuevo sistema de software. Más allá del tipo de dominio y del proceso de requisitos utilizado, cabe destacar que el UdeD es único. Suele ser separado para estudiar por un lado los procesos del negocio actual y por el otro la planificación de los procesos del negocio futuro. Esta división abstracta cumple con el objetivo de conocer el problema en estudio antes de definir una solución.
Como ya se mencionó, el proceso de requisitos [8] se encuentra entre el grupo de los procesos que estudian por un lado el dominio actual y luego analizan los procesos del negocio para cuando el nuevo sistema de software se encuentre en ejecución. Estos cambios quedan claramente reflejados en los escenarios, ya que los escenarios actuales evolucionan a los escenarios futuros. Pero no sucede lo mismo con el LEL, que se construye tempranamente en el proceso de requisitos y es utilizado como ancla en todos los modelos posteriores. La única actualización que sufre este glosario corresponde a mejoras relativas

referidas al vocabulario del dominio actual, por lo tanto la evolución *aparente* reflejada en los escenarios futuros no es acompañada en el LEL, generando confusión e incorporando ambigüedad en los modelos construidos.

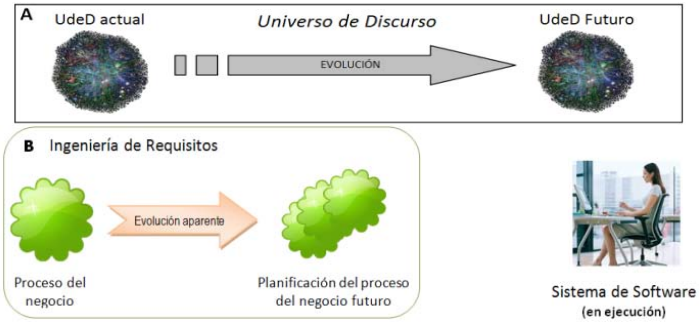


Figura 2 – Evolución vs. Evolución *aparente*

En la Figura 2 parte A, se ilustra la evolución del UdeD donde existe un UdeD actual y un UdeD futuro. El proceso de requisitos se desarrolla en su totalidad en el UdeD actual que tiene la particularidad de estar disponible desde el inicio del proceso de construcción del software. Por otro lado, el UdeD futuro se compone de los procesos del negocio futuro ya materializados en el nuevo sistema de software. Cabe destacar que este UdeD futuro en realidad comienza desde que se empieza a pensar en el nuevo sistema de software hasta pasado un tiempo después que fue implementado.
En la Figura 2 parte B se muestra la evolución *aparente* del UdeD, aquella evolución que se produce particularmente durante la Ingeniería de Requisitos. Esta evolución se lleva a cabo en un lapso relativamente corto de tiempo. Inicia con la necesidad del nuevo sistema de software y finaliza con la ERS. Esta evolución de los procesos del negocio es meramente conceptual ya que sólo se encuentra representada en los modelos de requisitos construidos.

3 Vocabulario de los Escenarios Futuros

La mera introducción de un cambio en un proceso de negocio puede generar una alteración en el vocabulario. Este impacto en el vocabulario puede ser mensurable en función de los términos que cambian de significado, de términos nuevos que aparecen y de términos que dejan de tener una relación directa con ese proceso de negocio que se está describiendo.
Un glosario que no evoluciona oculta la polisemia del vocabulario existente en cualquier contexto organizacional. Esta polisemia se da naturalmente por el dinamismo de los procesos del negocio y de manera particular y esperable durante un proceso de construcción del software, donde existe una decisión de modificar dichos procesos.
Es posible que la evolución del vocabulario sea el primer indicador de un cambio en el dominio (ver Figura 3). Esto se debe a que es necesario modificar el léxico para poder describir los planes de cambios en los procesos del negocio. El inicio anticipado de la evolución *aparente* del léxico se debe a la necesidad de contar con un vocabulario apropiado para describir las nuevas actividades, tareas, roles, recursos, etc. Los cambios en el vocabulario pueden preceder, ser simultáneos o suceder a los cambios en los procesos. Es así que el vocabulario puede evolucionar a partir de una modificación que se piensa para el

futuro sobre una tarea o proceso del negocio (ej. Es necesario controlar el ingreso y egreso del stock); puede evolucionar a partir de un cambio efectivizado en un proceso (ej. el ingreso como el egreso de artículos al stock se debe realizar con el formulario correspondiente); o evolucionar durante la Ingeniería de Requisitos, ya que el propio proceso induce a pensar los cambios.
Por otro lado, la estabilización del vocabulario es posterior a que el sistema se encuentre en ejecución (ver Figura 3), donde la evolución del LEL se ralentiza luego de que el sistema haya estado en uso por un período, cuando los usuarios adopten total o parcialmente dicho léxico creado durante la Ingeniería de Requisitos. Este nuevo vocabulario incluirá otro conjunto de términos adicionales a los creados para construir los escenarios futuros y la ERS. Dichas preferencias serán determinadas por la cultura personal de los clientes-usuarios (ej. el haber utilizado con anterioridad un software similar, tener experiencia en el área o en el tema en cuestión, etc.) y por la cultura organizacional (ej. las cosas se han llamado siempre de una manera determinada y a pesar que la evolución de los procesos modificó su significado, por costumbre se continúa utilizando la denominación anterior). Este será un glosario final que incorporarán los clientes-usuarios a sus actividades cotidianas y conformará un LEL evolucionado para un futuro proceso de requisitos.

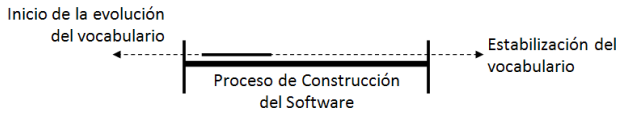


Figura 3 – Evolución del vocabulario

Particularmente para la Ingeniería de Requisitos, la evolución del vocabulario comienza en el momento de pensar en los cambios o en la solución hasta que dichos cambios se efectivizan en el diseño. Los cambios planificados para los procesos del negocio que no son registrados en el vocabulario, generan una desactualización rápida del mismo. El LEL cumple inicialmente con los objetivos previstos y un cubrimiento del 100% en las descripciones de los escenarios actuales. Al planificar la solución y describir los escenarios futuros, este cubrimiento disminuye aproximadamente en un 30%. Este valor puede variar según los cambios previstos en los procesos del negocio. El LEL_R ha sido aplicado desde 2010 en la carrera de grado Ingeniería en Informática de la UNLaM[2]. Los alumnos utilizaron el proceso [8] para crear una ERS. En el análisis de estos casos se ha detectado un cambio en el vocabulario entre el LEL y el LEL_R de aproximadamente un 30%.
Es de esperar que cuanto mayor sea el cambio en los procesos del negocio, más vocabulario se necesitará para describirlos. Es durante la Planificación del UdeD futuro donde se genera una mayor ampliación de la semántica de los términos existentes, la incorporación de términos nuevos y el desuso de algunos otros. Este es un fenómeno de la misma naturaleza al de la introducción de neologismos en un idioma. Un término nuevo puede ingresar y luego desvanecerse o perdurar volviéndose relevante. Esto hace necesario que el nuevo glosario registre, tanto como sea posible, este fenómeno. Este nuevo glosario es utilizado para mejorar la legibilidad y comprensión de los escenarios futuros y deberá ser levemente

[2] Son casos pseudo reales que los alumnos proponen para realizar los trabajos prácticos de la cátedra Ingeniería de Requerimientos.

actualizado para cubrir también la ERS ya que las transformaciones desde los escenarios futuros a los requisitos requerirán de algunas ampliaciones léxicas.
En este punto se abren varios caminos posibles ya que la decisión de cómo incorporar el nuevo vocabulario no es trivial. El resultado de una incorrecta documentación puede perjudicar seriamente la comprensión y legibilidad de los documentos generados, incluso de la propia ERS.

4 Estrategias de Construcción del LEL_R

En esta sección se analizan dos aspectos fundamentales para la construcción del LEL_R. Por un lado, la estructura del documento y por el otro, el enfoque de construcción.
Con respecto a la *estructura del documento* existen dos posibilidades. Una es modificar el LEL y describir todas las vistas del vocabulario en un único documento. La otra opción es dejar el LEL para representar el vocabulario del UdeD actual y generar un documento independiente para el nuevo glosario. La registración del vocabulario utilizado en el proceso de requisitos dentro del mismo LEL fue descartada debido a que estos glosarios tienen diferentes objetivos. Uno describe el vocabulario inicial del dominio, el otro el vocabulario necesario para planificar la solución. La inclusión en el LEL de los nuevos términos agrega complejidad dificultando no sólo su construcción sino también su posterior lectura. Esto se debe a la necesidad de utilizar gran cantidad de homónimos, para diferenciar las descripciones correspondientes al vocabulario del UdeD actual de aquellas necesarias para el proceso de requisitos. Por lo tanto, tener un único documento para ambos vocabularios, en vez de mejorar la calidad del LEL y reducir la ambigüedad de los documentos de requisitos, tiende a incorporar riesgos no deseados.
En este trabajo se ha optado por crear un nuevo glosario en un documento independiente denominado, como ya se ha mencionado, LEL_R. Este nuevo glosario comparte muchas de las particularidades del LEL y gran parte del proceso de construcción. Algunas diferencias entre ambos glosarios son las siguientes:

- El LEL describe el vocabulario del UdeD actual. El LEL_R describe el vocabulario de los documentos del proceso de requisitos.
- El LEL es un vocabulario vigente. El LEL_R es parcialmente artificial.
- El LEL se genera muy tempranamente en el proceso de requisitos. El LEL_R se crea con la Planificación del UdeD futuro.

Con respeto al *enfoque de construcción* del LEL_R puede ser realizado de manera interactiva o en batch.
La construcción interactiva del LEL_R implica realizar el proceso de construcción del LEL_R en paralelo con el proceso de construcción de los escenarios futuros. Durante la descripción de cada escenario futuro se analiza el vocabulario utilizado y se identifica el nuevo vocabulario o aquel que cambia de significado, ambos modelos se construyen en simultáneo. La ventaja de este enfoque es que en el momento de construir el glosario se cuenta con la información directa del UdeD. Pero este enfoque genera una sobrecarga de atención y concentración para el ingeniero de requisitos ya que debe construir dos modelos en paralelo que a pesar de ser complementarios tienen objetivos diferentes e inter-relaciones e intra-relaciones particulares para cada caso. Este exceso de objetivos puede disminuir la calidad de alguno de los dos modelos (escenarios futuros o LEL_R), llegar a reducir la calidad de ambos o afectar la calidad de los modelos que los utilizan. Por otro

lado, se dificulta cumplir con el principio de circularidad del LEL, donde se maximiza el uso de otros símbolos en las descripciones, ya que los escenarios futuros que contendrán ese nuevo vocabulario aún no se han descripto.
La construcción en batch se refiere a construir el LEL_R una vez finalizado el proceso de construcción de los escenarios futuros. Se debe analizar el vocabulario utilizado en cada escenario. En este momento toda la atención del ingeniero de requisitos se concentra en el vocabulario necesario para minimizar la ambigüedad de los escenarios futuros y en construir un glosario de buena calidad. Este enfoque permite una visión en dos etapas con mecanismos cognitivos distintos. En la primera etapa la atención está puesta en describir y definir el nuevo sistema de software. Mientras que en la segunda etapa de construcción del LEL_R, se está pensando particularmente en el vocabulario y no en los procesos del negocio.
La construcción del LEL_R en batch también tiene algunos inconvenientes con la particularidad que son mitigados con la mera generación de una lista inicial de símbolos. Unos de estos problemas es que los escenarios futuros se describen sin un vocabulario que los complemente o con el LEL que lo hace parcialmente. Esto puede provocar que los escenarios futuros requieran ser refinados luego de construir el nuevo glosario para eliminar detalle de los escenarios que ahora se encuentran en el LEL_R.
En resumen, la estrategia a la que adhiere el presente artículo consiste en realizar un documento independiente para el LEL_R utilizando un enfoque de construcción en batch.

5 Heurística de Construcción del LEL_R

Esta heurística se inicia generando sólo una lista de símbolos candidatos en paralelo con la primera versión de los escenarios futuros. Dicha lista se completa durante las entrevistas donde se negocian y se validan estos escenarios. Tanto de los casos realizados como de la mera observación de la actividad a realizar surge que la mayor actividad de la construcción del LEL_R se concentra una vez que los escenarios futuros están totalmente construidos. Es probable que durante la generación de la ERS pueda aparecer nuevo vocabulario o se modifique alguna noción o impacto de un símbolo ya existente en el LEL_R. Se estima que estos cambios serán de muy baja incidencia. A continuación se describe la heurística de construcción del LEL_R:

1. IDENTIFICAR SIMBOLOS

1.1. Los escenarios futuros utilizan, en primera instancia, al LEL como referencia léxica, pero se debe prestar una especial atención a la aparición de nuevos términos necesarios para describir los cambios en los procesos del negocio. Estos símbolos nuevos deben ser identificados (marcados en el escenario) de alguna manera que los diferencie de los símbolos del LEL.

1.2. Al describir un escenario futuro también puede ser percibido un mal uso o el cambio de significado de un símbolo del LEL, en este caso se puede identificar, por ejemplo, con una "(h)" al final del símbolo para indicar que se está en presencia de un homónimo.

1.3. **CREAR LA LISTA DE SIMBOLOS CANDIDATOS FUTUROS**

1.4. Durante las entrevistas para negociar y validar los escenarios futuros se debe generar una lista denominada Lista de Símbolos Candidatos Futuros, con las palabras o frases que el usuario utiliza para describir sus necesidades, los nuevos procesos o sus deseos y que no son parte del vocabulario utilizado hasta el momento, por lo tanto no se encuentran en el LEL.

1.5. Nuevamente, si durante las entrevistas se percibe un mal uso o el cambio de significado de un símbolo del LEL identificarlo como homónimo.
1.6. Clasificar los símbolos en Sujeto, Objeto, Verbo y Estado.

2. **CREAR LEL_R**
   2.1. Al finalizar con la descripción de los escenarios futuros se debe crear el LEL_R e incorporarle los símbolos de la Lista de Símbolos Candidatos Futuros.
3. **DESCRIBIR SIMBOLOS**
   3.1. Describir los símbolos del LEL_R retornando al UdeD para completar su definición. Utilizar los patrones del LEL para los tipos Sujetos, Objetos, Verbos y Estados.
   3.2. Recorrer cada escenario futuro:
      3.2.1. Verificar el cubrimiento del LEL:
         3.2.1.1. Por cada símbolo del LEL utilizado en el escenario futuro, ir a su definición original. Si el cubrimiento del LEL es del 100% copiarlo al LEL_R identificado como migrado, o sea que mantiene su significado durante la planificación del UdeD futuro.
         3.2.1.2. Si parte de la noción o del impacto no corresponde, crear un símbolo nuevo con la descripción correcta e identificarlo como homónimo.
      3.2.2. Por cada símbolo identificado como nuevo:
         3.2.2.1. Verificar si ya existe en el LEL_R. Si existe agregar el sinónimo. Si no existe, agregar el símbolo y no utilizar ningún identificador.
      3.2.3. Por cada símbolo identificado como homónimo.
         3.2.3.1. Verificar si ya existe en el LEL_R. Si existe agregar el sinónimo. Si no existe, agregar el símbolo e identificarlo como homónimo.
         3.2.3.2. Se debe controlar si este nuevo símbolo reemplaza la definición original del LEL o es necesario tener ambos. En este último caso copiar el símbolo original en el LEL_R.
4. **VERIFICAR LEL_R**
   4.1. Con el LEL_R completo utilizar el proceso de inspección del LEL [15] para verificar su consistencia interna. Este proceso deberá ser mínimamente adaptado. Se debe verificar también parte de su consistencia externa, controlando que todos los escenarios futuros utilizados existan en el conjunto de escenarios futuros.

Se debe controlar que el LEL_R cumpla con el principio de Vocabulario Mínimo y el principio de Circularidad. Este último sufre algunas modificaciones. En este glosario se reduce el uso de Verbos. Los impactos que deben hacer referencia a un verbo lo hacen ahora a un escenario futuro (ver Figura 5). Si existe un escenario futuro que responda a la acción que se desea describir se vincula al mismo. Si no existe se puede crear un símbolo Verbo pero se sugiere revisar previamente los escenarios futuros. Estos impactos de un Verbo deben ser identificados escribiendo el título del escenario en mayúscula. Con los Sujetos es diferente, ya que el uso de un escenario futuro como impacto puede estar representando acciones de diferentes actores, excediendo la información necesaria para el impacto. Se sugiere referenciar al escenario como en los otros casos, pero su lectura se debe reducir sólo a aquellos episodios donde el Sujeto en cuestión tiene un rol determinado, o sea es un actor del episodio.
El símbolo Sistema no debe ser descripto ya que en sus impactos estarán gran parte de los requisitos del software. Esto puede traer confusión al lector y desviar la atención del

objetivo de este modelo que es describir vocabulario. Se debe utilizar la palabra "Sistema" con el rol Sujeto cuando sea necesario pero no identificarlo como un símbolo.
Debido a que este glosario cuenta con símbolos artificiales y que se ha utilizado el concepto de homónimo para evidenciar la evolución *aparente*, se sugiere reducir tanto como se pueda el uso homónimos entre nuevos símbolos. En caso de ocurrir, una posibilidad es buscar otro nombre para alguno de los símbolos en cuestión.

La Figura 4 muestra un resumen de los pasos que se deben seguir para construir el LEL_R y en la Figura 5 un ejemplo de un símbolo Sujeto del LEL que se modificó al definir la Planificación del UdeD futuro. En el símbolo del LEL se puede observar que los últimos tres impactos son símbolos verbos. En cambio los impactos del LEL_R son escenarios futuros y en dichos escenarios el nuevo sistema de software tiene un rol principal en los episodios.

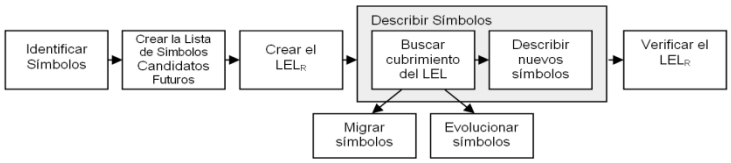


Figura 4 – Resumen de la heurística de construcción del LEL_R

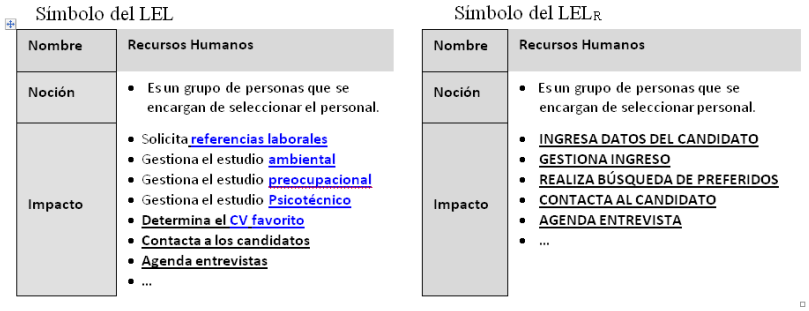


Figura 5 - Ejemplo de un símbolo Sujeto y de su Evolución *aparente*

6 Conclusiones

En el presente artículo se analizó el impacto de la evolución *aparente* lingüística durante los procesos de requisitos en general y en particular en el proceso de requisitos basado en escenarios [8]. Los cambios en el léxico comienzan a generarse desde el momento en que se presenta la necesidad de cambiar un proceso de negocio y finaliza bastante después de la puesta en servicio del sistema.
La incorporación en el dominio de los cambios en el vocabulario generados durante el proceso de requisitos, puede ser parcial o total y necesitar mucho más tiempo que el propio desarrollo de software. A pesar de esto, a la hora de asegurar los requisitos del software, es necesario contar con un vocabulario comprensible por todos los involucrados y respaldado por un documento propio del proceso.
La evolución *aparente* del vocabulario puede quedar oculta en el uso de herramientas, métodos o en los mismos procesos. Ignorar los cambios en el vocabulario y dejar vigente

un glosario que se desactualizó, puede llevar a comprender erróneamente los requisitos del software. Esta es una problemática que debe ser atendida y registrada en el momento correcto, con el objetivo de evitar confusiones y errores al momento de definir la solución.
Aproximadamente la diferencia entre el LEL y el LEL_R es del 30% (como se mencionó en sección 3) de los símbolos y están esencialmente relacionados con aquellas actividades de las personas que son tomadas por el nuevo sistemas de software.
La creación del LEL_R ha observado una mejora significativa en la comprensión y calidad de los escenarios futuros, lo que ha generado a su vez una mejora en la calidad de la ERS. Se están elaborando métricas de calidad para ponderar apropiadamente estas mejoras.
En los numerosos casos de estudio realizados, inicialmente se aplicó una guía básica y con los resultados obtenidos se generó la heurística presentada, la cual fue aplicada en más casos. Se espera continuar con las pruebas y profundizar aún más en la heurística analizando aspectos de trazabilidad y el uso de alguna herramienta.

7 Bibliografía

[1] Rolland C., Ben Achour C. (1998) Guiding the construction of textual use case specifications, Data & Knowledge Engineering 25, pp 125-160.

[2] Oberg R., Probasco L, Ericsson M. (1998) Applying Requirements Management with Use Cases. Rational Software Corporation.

[3] Weidenhaupt K., Pohl K., Jarke M., Haumer, P (1998) Scenarios in System Development: Current Practice., IEEE Software, pp 34-45.

[4] Alpaugh T.A., Antón A.I., Barnes T., Mott B.W. (1999) An Integrated Scenario Management Strategy, International Symposium On Requirements Engineering (RE99), Limerick-Irlanda (IEEE Computer Society Press), 142-149.

[5] Robertson S. and Robertson J. (2006) Mastering the Requirements Process, 2nd Ed, Addison-Wesley.

[6] Leite, J.C.S.P., "Engenharia de Requisitos", Notas Tutoriales, material de enseñanza en el curso Requirements Engineering, Computer Science Department of PUC-Rio, Brasil, 1994.

[7] Brietman K, Liete J., Berry D. (2004) Supporting Software Evolution, Requirements Engineering, May 2005, Volume 10, Issue 2, pp 112-131

[8] Leite J.C.S.P., Doorn J.H., Kaplan G.N., Hadad G.D.S., Ridao M.N. (2004) Defining System Context using Scenarios, In: Leite J.C.S.P. and Doorn J.H (eds) Perspectives on Software Requirements, Kluwer Academic Publishers, ch. 8, pp.169-199.

[9] Leite J.C.S.P., Franco, A.P.M., (1990) "O Uso de Hipertexto na Elicitaçao de Linguagens da Aplicaçao", Anais de IV Simpósio Brasilero de Engenharia de Software, SBC, pp. 134-149.

[11] Kaplan G.N., Doorn J.H., Hadad G.D.S. (2008) Handling Extemporaneous Information in Requirements Engineering, Encyclopedia of Information Science and Technology, editor: Mehdi Khosrow- Pour, D.B.A., Information Science Reference, EEUU, ISBN: 978-1-60566-026-4, 2º edición, pp.1718-1722.

[12] Doorn J., Hadad G., Kaplan G. (2002) Comprendiendo el Universo de Discurso Futuro, WER'02 - Workshop on Requirements Engineering, Valencia, Spain, pp.117-131.

[13] Hadad G.D.S., Doorn J.H., Kaplan G.N. (2008) Creating Software System Context Glossaries, In: Mehdi Khosrow-Pour (ed) Encyclopedia of Information Science and Technology. IGI Global, Information Science Reference, Hershey, PA, USA, ISBN: 978-1-60566-026-4, 2nd edn, Vol. II, pp. 789-794.

[14] Leite J.C.S.P., Hadad G.D.S., Doorn J.H., Kaplan G.N. (2000). A Scenario Construction Process, Requirements Engineering Journal, 5, (1). 38-61.

[15] Kaplan, G.N., Hadad, G.D.S., Doorn, J.H., Leite, J.C.S.P., "Inspección del Léxico Extendido del Lenguaje", proceedings of WER'00 – III Workshop de Engenharia de Requisitos, Río de Janeiro, Brazil, 2000.

Generación de un Algoritmo de Ranking para Documentos Científicos del Área de las Ciencias de la Computación

H. Kuna[1], M. Rey[1], J. Cortes[1], E. Martini[1], L. Solonezen[1], R. Sueldo[1]

[1]Depto. de Informática, Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones.
{hdkuna, m.rey00}@gmail.com

Resumen. La generación de un algoritmo de ranking para el ordenamiento de documentos científicos pertenecientes al área de ciencias de la computación es un requerimiento fundamental para el desarrollo de Sistemas de Recuperación de Información que sean capaces de operar sobre tal tipo de elementos. Estos sistemas buscan optimizar el proceso de búsqueda de contenido en la web a través de diversas herramientas, entre ellas los meta-buscadores. Los mismos amplían el espectro de cobertura en la búsqueda, a partir de la capacidad para utilizar las bases de datos de varios buscadores en simultáneo; además de poder incorporar diversos métodos para el ordenamiento de los documentos, que mejoren la relevancia de los resultados para el usuario. En este trabajo se presenta el desarrollo de un algoritmo de ranking para ordenar el listado de resultados que retorne un Sistema de Recuperación de Información para la búsqueda de documentos científicos en el área de las ciencias de la computación.

Palabras clave: recuperación de información, algoritmo de ranking, búsqueda web, indicadores bibliométricos.

1 Introducción

1.1 Sistemas de Recuperación de Información

Un Sistema de Recuperación de Información (SRI) se puede definir como un proceso capaz de almacenar, recuperar y mantener información [1], [2]. Existen en la literatura diversas propuestas sobre la estructura básica que debiera tener un SRI, un ejemplo es la que lo considera a partir de la unión de cuatro elementos como son [3]:

- Los documentos que forman parte de la colección sobre la que se realizará la recuperación.
- Las consultas que representan las necesidades de información por parte de los usuarios.

- La forma en la que la modelan las representaciones de los documentos, consultas y las relaciones presentes entre ellos.
- La función de evaluación que determina para cada consulta y documento el orden que ocupará en los resultados a presentar.

En la actualidad los principales modelos de SRI que operan sobre internet son: los directorios, los buscadores y los meta-buscadores [4]. Considerando tal clasificación se puede afirmar que existen diversas implementaciones de SRI en la web que utilizan diferentes métodos de búsqueda sobre contextos generales o particulares, como se puede observar en distintas publicaciones [5], [6].

1.2 SRI para Documentos Científicos del Área de Ciencias de la Computación

No se ha encontrado evidencia de la existencia de implementaciones de SRI que sean aplicadas específicamente a bases de datos de documentos científicos pertenecientes al área de ciencias de la computación, que además implementen diversos métodos para la mejora del listado de resultados a presentar al usuario en base a la relevancia que puedan tener los mismos con respecto a la consulta efectuada.

En el contexto del presente trabajo cobran una mayor notoriedad los meta-buscadores, debido a que posibilitan la utilización de bases de datos de otros buscadores, replicando las consultas de los usuarios sobre cada una de ellas y, posteriormente, procesar los resultados obtenidos de la manera que se crea conveniente para generar un único listado de resultados a presentar al usuario.

La generación de un SRI que opere sobre documentos científicos del área de ciencias de la computación, requiere directamente el desarrollo de diversos componentes, entre los cuales se destaca el algoritmo a utilizar para la evaluación de cada resultado obtenido de las búsquedas con el objetivo de fusionar y ordenar el listado final de resultados [5].

1.3 Métricas para la Evaluación de Documentos Científicos

Dada la naturaleza del SRI planteado y el algoritmo de ranking a generar para el mismo, los métodos para la evaluación de los resultados deben ser desarrollados en forma particular. Para la evaluación de documentos científicos se debe considerar una serie de características evaluables, como ser [7], [8]:

- El tipo de fuente de publicación, distinguiendo si el mismo se publica en una revista científica o en un congreso científico o evento similar.
- La calidad de los autores, considerando en este caso la cantidad de publicaciones que ha realizado el mismo y la relevancia de las mismas, medida a través de la cantidad de citas que hubieran generado.
- La calidad del artículo en sí, en este caso, medida a través de la cantidad de veces que haya sido citado a lo largo del tiempo.

Para cada una de estas características existen métricas ampliamente aceptadas que pueden aplicarse, algunas de ellas pueden observarse con claridad en la tabla 1.

Table 1. Métricas relevadas para la evaluación de artículos científicos

Característica a evaluar	Métricas disponibles		Origen de la métrica
Tipo de fuente de publicación	Publicación en Revista Científica	Factor de Impacto (IF) [9]	Web of Knowledge[1] – Institute for Scientific Information (ISI)
		SCImago Journal Rank (SJR) [10]	Scopus[2] – Grupo SCImago, Univ. De Extremadura, España
	Publicación en Congreso Científico	Ranking CORE [11]	Computer Research & Education of Australia[3]
Calidad de los autores	Índice H [12]		Artículo científico
	Índice G [13]		Artículo científico
Calidad del artículo	Índice AR [14]		Artículo científico
	Cantidad de citas		-

En el caso del tipo de fuente de publicación, para aquellas publicaciones realizadas en revistas existen dos índices que se utilizan para estimar su calidad: por un lado el Factor de Impacto (IF, por sus siglas en inglés) [9]; y el índice SJR, SCImago Journal Rank [10]. En ambos casos se trata de métricas que toman las citas que reciben los artículos publicados en una revista y las evalúan tanto en cantidad como en lo referente a la relevancia que tiene la producción que la realiza. Mientras que en caso de que la publicación se realice en un congreso o evento similar existe un ranking como es el que genera en la web de la Computer Research & Education of Australia (CORE) [11] en donde a diversas conferencias o congresos se los ubica en uno de los cuatro niveles que tiene establecidos: A*, A, B y C, listado que se establece en la mencionada web que será reformado y actualizado a la brevedad.

Para estimar la calidad de la producción de un autor se dispone de métricas como pueden ser: el índice H [12] y el índice G [13]; lo que hacen éstas es tomar la cantidad de citas recibidas por las diferentes publicaciones del autor y la cantidad de publicaciones para calcular un valor que representa la influencia del mismo.

Para evaluar la calidad de una colección de publicaciones a través del tiempo se puede utilizar un índice como es el AR [14], que toma la antigüedad de las mismas y las pondera utilizando ese factor en combinación con la cantidad de citas obtenidas por cada uno de los artículos que componen la colección; siendo este último factor

[1] www.wokinfo.com – Accedido: 16/07/13
[2] www.scopus.com – Accedido: 16/07/13
[3] www.core.edu.au – Accedido: 16/07/13

otra de las métricas disponibles para evaluar la calidad de un documento científico particular.

El objetivo del presente trabajo es el de desarrollar un algoritmo de ranking para documentos científicos específico para el área de ciencias de la computación, de manera de generar un componente que pudiera ser incluido en un SRI, puntualmente un meta-buscador, cuyo propósito sea la recuperación de contenidos de esta área de conocimiento en la web. Para tal tarea se pretende incluir en el algoritmo diversas métricas que permiten la evaluación de un artículo científico desde diversos aspectos como son: el tipo de fuente de publicación, la calidad de los autores que lo suscriben y la calidad del artículo en sí.

2 Materiales y Métodos

2.1 Estructura del SRI

Dado que el algoritmo de ranking se debe incorporar a un SRI, se debió considerar cuál sería la estructura del mismo, estando la misma conformada por módulos para realizar las siguientes operaciones, un esquema de los mismos puede observarse en la figura 1:

- Tratamiento de las consultas introducidas por el usuario para ser utilizadas sobre las fuentes de datos integradas;
- Realización de la búsqueda replicando la consulta del usuario en las diferentes fuentes de datos;
- Captura, selección y unificación de los resultados obtenidos de las diferentes fuentes, siendo éste el módulo en el que el algoritmo a desarrollar se incorporaría;
- Mejoramiento de los resultados a presentar al usuario a través de diversas técnicas inteligentes.

Una cuestión más a considerar fueron las fuentes de datos a las que accedería el SRI para obtener los documentos científicos que coincidieran con los requerimientos del usuario, considerando que las mismas deberían permitir la obtención, directa o indirectamente, de los valores correspondientes a las métricas que se determinara utilizar en el algoritmo de evaluación. Los buscadores seleccionados inicialmente fueron: el buscador académico de Google, Google Scholar[4], y el buscador Scopus de la editorial Elsevier. Esta selección se debió a que ambas herramientas se encuentran en constante actualización tanto en sus componentes como en los documentos a los que acceden y cumplen con el requisito de disponer de diversas métricas que se pueden utilizar al evaluar a las publicaciones que recuperan, constituyendo alternativas de gran calidad para dar cumplimiento a los objetivos planteados [15], [16].

Con la definición del contexto en el cual operaría el algoritmo de ranking se prosiguió con el diseño, desarrollo y posterior validación del mismo.

[4] www.scholar.google.com – Accedido: 16/07/13

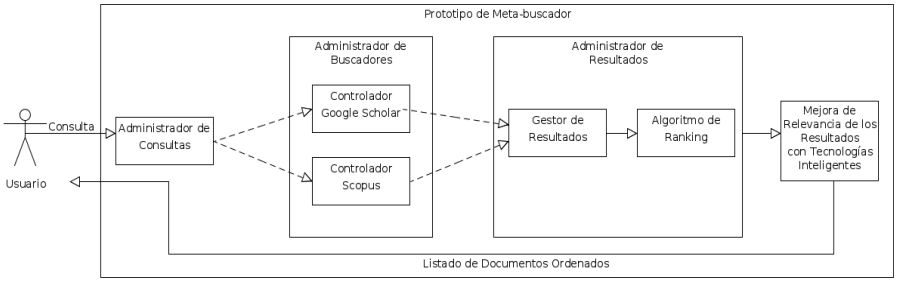


Figura 1. Componentes del meta-buscador

2.2 Diseño del Algoritmo

En una primera instancia se seleccionaron las métricas a utilizar dentro del algoritmo de ranking, priorizando diversos aspectos en cada una. Considerando nuevamente las características evaluables de los documentos científicos, la fuente de publicación, la calidad de sus autores y la calidad del artículo en sí; se determinaron las métricas a considerar para ponderar cada resultado:

- Para el tipo de fuente de publicación: se consideraron dos factores para la valoración de este punto, dependiendo si el artículo se publicó en una revista científica o en un congreso del área de conocimiento correspondiente. Para el primer caso, se ha optado por utilizar el índice SJR [10] desarrollado por el grupo de investigación SCImago; esta selección se debió a las ventajas que presenta con respecto al IF de ISI [9], como ser [17], [18]: es de acceso abierto; en la base de datos de Scopus contiene una mayor cantidad de revistas, incluyendo aquellas que no están escritas en inglés; no sólo hace una evaluación cuantitativa de las citas recibidas por un artículo sino que también lo hace en forma cualitativa, incorporando la calidad de la revista que genera la cita; entre otras. Para el caso de los artículos procedentes de congresos o reuniones científicas se empleó el ranking generado por la Computing Research and Education Association of Australia (CORE). En resumen:

  – *Si* (tipo_publicación = revista_científica) *Entonces* usar_SJR
  – *Si* (tipo_publicación = congreso_científico) *Entonces* usar_CORE

- Para la calidad de los autores: en este caso se optó por utilizar el índice H [12], aun considerando algunas críticas que puedan realizarse sobre el mismo, ya que es ampliamente aceptado y utilizado para la evaluación de la producción científica de un determinado autor [7], [8]. El índice concretamente representa la cantidad *X* de artículos de un autor que han recibido *X* citas como mínimo.
- Para la calidad del artículo: en este caso se determinó considerar ambas métricas relevadas previamente, el índice AR [14] y la cantidad de citas recibidas por el artículo, remarcando la necesidad de la primera de ser adaptada para trabajar sobre un único documento en vez de una colección como se plantea originalmente.

2.3 Desarrollo del Algoritmo de Ranking

Una vez seleccionadas las métricas que conformarían el algoritmo se procedió con el desarrollo concreto del mismo. Para tal actividad se definieron inicialmente las fórmulas a través de las cuales se calcularían los valores correspondientes para cada propiedad de los documentos:

- Para el factor correspondiente a la propiedad del tipo de fuente de publicación: en caso de que se trate de una publicación en una revista científica se calcula el logaritmo en base 10 del valor del índice SJR de la revista, esto con la finalidad de homogeneizar los valores de este factor con respecto al resto de los componentes del algoritmo, ya que el rango de valores presentes en el índice es mayor a la decena en un gran número de revistas. Mientras que para el caso de que la publicación se realice en un congreso o evento científico se debió adaptar el modelo de clasificación que otorga el ranking CORE, transformando a un formato numérico la clasificación del congreso para poder operar con él. El valor correspondiente al factor de la fuente de publicación se obtiene mediante la fórmula 1, en caso de haber sido en una revista, y con la fórmula 2, en caso de haber sido en un congreso.

$$\text{fuentePublicacion} = \log10(\text{SJR}) \; . \tag{1}$$

$$\text{fuentePublicacion} = [A^* = 1;\ A = 0.75;\ B = 0.5;\ C = 0.25] \; . \tag{2}$$

- Para el factor correspondiente a la calidad de los autores: se considera el índice H del autor del artículo en evaluación. En caso de que se trate de un artículo con más de un autor se pondera el valor del índice con respecto a la posición que ocupa en el listado de autores del documento. Además se vuelve a utilizar el logaritmo en base 10 para el valor resultante de la sumatoria ponderada de los valores del índice H de los autores del artículo. El cálculo del factor se puede ver en la fórmula 3.

$$\text{autores} = \log10(\sum(\text{indiceH}(\text{autor}_i))/i) \; . \tag{3}$$

- Para el factor correspondiente a la calidad del documento en evaluación: en este caso, dado en enfoque combinado al emplear como base al índice AR y anexar al mismo la cantidad de citas recibidas por la publicación, se determinó que la el factor ponderaría la calidad de la misma a través del cociente resultante entre ambos elementos: la antigüedad y la cantidad de citas. Dando origen a la fórmula 4 en la que se puede observar el resultado de la adaptación realizada.

$$\text{calidadPublicacion} = \text{citasRecibidas} \,/\, \text{antigüedadPublicacion} \; . \tag{4}$$

Una vez determinados los componentes correspondientes a cada una de las características a evaluar de un documento científico, se determinó que se adicionaría al cálculo final del algoritmo un factor de ajuste, el cual tendría la función de permitir que uno de los factores tuviera más importancia que los otros. El cálculo de los factores asociados a cada propiedad multiplicados por los factores de ajuste resulta en el valor final que se utiliza para realizar el orden de los resultados antes de presentarlos al usuario.

- Con la inclusión de los factores de ajuste asociados a los componentes correspondientes a las propiedades evaluadas se da forma al valor final correspondiente a cada documento en evaluación por parte del algoritmo de ranking, esto se refleja en la fórmula 5. Los valores establecidos, en forma conjunta con los expertos en la temática, para los factores de ajuste fueron: 0.5, 0.3 y 0.2 respectivamente.

$$valorFinal= \alpha*[fuentePublicacion] + \beta*[autores] + \gamma*[calidadPublicacion] . \quad (5)$$

3 Experimentación

3.1 Desarrollo del Prototipo de SRI para la Experimentación

Con el objetivo de validar el correcto funcionamiento del algoritmo de ranking propuesto se ha incluido al mismo dentro de un prototipo de meta-buscador, el cual constituye la implementación parcial del SRI descripto en las secciones anteriores.

El prototipo mencionado fue desarrollado priorizando el uso de tecnologías que fueran basadas en la filosofía Open Source, como ser: los lenguajes HTML, PHP y SQL, junto al motor de bases de datos MySQL, utilizando como entorno para su implementación al servidor web Apache.

El proceso de implementación del prototipo se descompuso en los siguientes pasos:

1. Desarrollo de los métodos para acceder, consultar y extraer los resultados de los buscadores Google Scholar y Scopus.
2. Implementación del algoritmo de ranking con el acceso a las fuentes de datos que almacenan los valores de las diferentes métricas involucradas.
3. Desarrollo de los componentes visuales del prototipo, es decir, de las interfaces para captura de las consultas del usuario y la correspondiente a la presentación del listado de resultados unificado.
4. Integración de todos los componentes en un único producto software.

3.2 Validación del Algoritmo Desarrollado

El proceso de validación constó de dos etapas que evaluaron los resultados desde dos perspectivas, inicialmente se ha considerado a los resultados desde la óptica de un experto en bibliotecología y posteriormente se ha evaluado al algoritmo de ranking

como componente del prototipo de SRI encargado de la mejora de la relevancia de los resultados a presentar al usuario final, para lo cual se ha contado con la colaboración de tres expertos en la temática de desarrollo de métodos de recuperación de información a partir de la web.

Para la primera instancia de validación, cuyo detalle puede observarse en la tabla 2, se han realizado diversas consultas, utilizando el prototipo de meta-buscador descripto en la sección anterior, operando sobre un número reducido de documentos, y exportando los resultados de los cálculos correspondientes al algoritmo de ranking a un archivo externo al SRI. Considerando tales datos el experto en el área de bibliotecología, ha determinado que las métricas empleadas han sido calculadas en forma correcta, generando un valor numérico que permite establecer un orden entre los documentos, que forman parte del listado resultante de la búsqueda, en base a la relevancia de los mismos evaluada a partir de las propiedades seleccionadas.

Table 2. Resultados de la primera instancia de validación

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por el experto
data mining AND outliers	20 (10 Google Scholar + 10 Scopus)	74%
fuzzy sets AND clustering	20 (10 Google Scholar + 10 Scopus)	87%
alphanumeric data AND outliers	20 (10 Google Scholar + 10 Scopus)	81%
scientific production AND metrics	20 (10 Google Scholar + 10 Scopus)	77%
text mining AND ontologies	20 (10 Google Scholar + 10 Scopus)	96%

Posteriormente se procedió con la valoración del algoritmo de ranking como componente del prototipo de SRI por parte de los expertos en la temática, el detalle de la experimentación se observa en la tabla 3, en la que se incrementaron la cantidad de consultas y la cantidad de resultados a obtener. En este caso el modo de trabajo de los expertos consistió en la evaluación de la calidad de los resultados con respecto a los diversos requerimientos del usuario. Como resultado, se ha determinado que el componente de gestión de los resultados, a través del algoritmo de ranking desarrollado, cumple satisfactoriamente con el objetivo de evaluar la calidad de los resultados para la generación del listado final a presentar al usuario, logrando que el mismo presente en sus primeros lugares a aquellos documentos científicos de mayor calidad.

Table 3. Resultados de la segunda instancia de validación

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por los expertos
data mining AND outliers	100 (50 Google Scholar + 50 Scopus)	72%
fuzzy sets AND clustering	100 (50 Google Scholar + 50 Scopus)	93%
alphanumeric data AND outliers	100 (50 Google Scholar + 50 Scopus)	84%
scientific production AND metrics	100 (50 Google Scholar + 50 Scopus)	90%
text mining AND ontologies	100 (50 Google Scholar + 50 Scopus)	94%
data mining AND systems audit	100 (50 Google Scholar + 50 Scopus)	69%
scientific articles AND ranking algorithms	100 (50 Google Scholar + 50 Scopus)	83%
fuzzy controllers AND robotics	100 (50 Google Scholar + 50 Scopus)	79%
fuzzy sets AND document processing	100 (50 Google Scholar + 50 Scopus)	75%
web agents AND document analysis	100 (50 Google Scholar + 50 Scopus)	86%

4 Conclusiones y Trabajos Futuros

Con el presente trabajo se ha conseguido desarrollar y validar un algoritmo de ranking específico para la evaluación de documentos científicos pertenecientes al área de ciencias de la computación. Para el mismo se han tomado en consideración distintos indicadores bibliométricos, con la finalidad de obtener valores para la evaluación de las distintas propiedades a fin de determinar la calidad de documentos científicos del área de ciencias de la computación. Además se ha incluido al mismo dentro de un prototipo de SRI, concretamente un meta-buscador, cuyo campo de aplicación son los documentos antes mencionados, constituyendo un avance significativo en lo que respecta a los objetivos del presente trabajo.

Como trabajos a futuro se pueden mencionar: evaluar la incorporación de otros indicadores bibliométricos que puedan ser de utilidad para el algoritmo de ranking, considerando en todo momento la especificidad propia del área a la que pertenecen

los documentos a evaluar; evaluar la incorporación de elementos propios de la lógica difusa y/o inteligencia artificial para automatizar adaptación de los factores de ajuste del algoritmo de ranking; incorporar al análisis de cada artículo resultante una evaluación de la reputación de sus autores para la sub área temática sobre la que se realice la búsqueda; entre otros.

5 Bibliografía

1. Salton, G., Mcgill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., (1983)
2. Kowalski, G.: Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers, Norwell, MA, USA (1997).
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. ACM press, New York (1999)
4. Olivas, J. A.: Búsqueda Eficaz de Información en la Web. Editorial de la Universidad Nacional de La Plata (EDUNLP), La Plata, Buenos Aires, Argentina (2011)
5. Serrano-Guerrero, J., Romero, F. P., Olivas, J. A., de la Mata, J.: BUDI: Architecture for fuzzy search in documental repositories. Mathw. Soft Comput., 16, 1, 71–85 (2009)
6. de la Mata, J., Olivas, J. A., Serrano-Guerrero, J.: Overview of an Agent Based Search Engine Architecture, en Proc. Of the Int. Conf. On Artificial Intelligence IC-AI'04, 62-67. Las Vegas, USA (2004)
7. Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R.: A Principal Component Analysis of 39 Scientific Impact Measures. Plos One, (2009)
8. Pendlebury, D. A.: The use and misuse of journal metrics and other citation indicators. Arch. Immunol. Ther. Exp. (Warsz.), 57(1), 1-11 (2009)
9. Garfield, E.: The history and meaning of the journal impact factor. JAMA, 295(1), 90-93 (2006)
10. Gonzalez-Pereira, B., Guerrero-Bote, V., Moya-Anegon, F.: The SJR indicator: A new indicator of journals' scientific prestige, arXiv:0912.4141, (2009)
11. CORE Conference Ranking, Computer Research & Education of Australia, http://www.core.edu.au
12. Hirsch, J. E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A., 102(46), 16569-16572 (2005)
13. Egghe, L.: Theory and practise of the g-index. Scientometrics. 69(1), 131-152 (2006)
14. Jin, B.: The AR-index: complementing the h-index. Issi Newsl. 3(1), p. 6 (2007)
15. Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., Herrero-Solana, V.: Coverage analysis of Scopus: A journal metric approach. Scientometrics. 73(1), 53-78 (2007)
16. Meho, L. I., Yang, K.: A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar. arXiv e-print cs/0612132, (2006)
17. Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., Karageorgopoulos, D. E.: Comparison of SCImago journal rank indicator with journal impact factor. Faseb J. 22(8), 2623-2628 (2008)
18. Leydesdorff, L., Moya-Anegón, F., Guerrero-Bote, V. P.: Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. J. Am. Soc. Inf. Sci. Technol., 61(2), 352–369 (2010)

Gestión de Contenido Organizacional (ECM)

Alejandra López, Julio Moreyra

Facultad de Ingeniería – Sede Trelew, Universidad Nacional de la Patagonia San Juan Bosco
Belgrano y Roca – Trelew – Chubut – Argentina
Malopez77@gmail.com, Julio.moreyra@gmail.com

Abstract. Enterprise Content Management makes reference to the handling of organizational content through printed or electronic means. It comprises the strategies, methods, and tools used to capture, handle, protect, preserve, and submit content and documents related to organizational processes.

The ECM is employed by several organizations nowadays since it provides help in the organizational process and thanks to this, it increases the company's profitability by means of improvement of the aspects related with the organization and management of documentation.

We studied the general aspects of ECM and three different products are analyzed: Alfresco, Nuxeo, and SharePoint as a ground for the comprehension of this methodology, its characteristics, scopes, benefits, and suitability.

A comparison is drawn between them based on the different perspectives: cost, community support, technology, options of execution environment, etc.

Keywords: Enterprise Content Management. Information lifecycle.Workflows.

1 Introducción

La gestión de contenido organizacional (ECM), está constituida por estrategias, métodos y herramientas usadas para capturar, gestionar, almacenar, preservar y distribuir contenido y documentos relacionados con los procesos organizacionales. Las herramientas ECM y sus estrategias permiten la gestión de la información no estructurada de la organización, donde quiera que la información se encuentre almacenada. Además, esta información debe poder ser utilizado de forma tal de alcanzar los objetivos de la organización. Son centrales a esta estrategia las herramientas y tecnologías ECM, que manejan el ciclo de vida completo de la información.

Hay cuatro áreas primarias, en las que ECM y los contenidos son fundamentales para el éxito de la organización: Cumplimiento de normas, Colaboración, Continuidad y Costos.

Cumplimiento de normas: Existen en EE.UU. regulaciones como Sarbanes-Oxley[1], HIPAA[2], DoD 5015.2[3] entre otros, que establecen formatos y requerimientos para la gestión de registros electrónicos de documentación. En nuestro país no existen este marco regulatorio.

Debido a normativas cada vez más estrictas y con el objetivo de mostrar niveles de transparencia más altos, las organizaciones de todo el mundo recurren a tecnologías y software de gestión de los archivos electrónicos. Los costos asociados para cumplimentar una regulación, como Sarbanes-Oxley o HIPAA, son mayores e imponen nuevas exigencias en la gestión de registros. Para ayudar a limitar los riesgos y costos deben desarrollarse estrategias ECM proactivas dentro de áreas claves, tales como gestión de registros y procesos de negocios, asegurando que las prácticas apropiadas del negocio sean seguidas y el contenido sea convenientemente capturado, almacenado, gestionado y descartado en el momento oportuno y legal dentro de su ciclo de vida. Las herramientas ECM, cuando se utilizan correctamente, pueden ayudar a reducir los costos generales en el cumplimiento de normas de la organización.

Colaboración: esta permite que las personas con áreas de conocimiento complementarias o solapadas puedan crear mejores resultados de forma más expeditiva que en el pasado. Con las herramientas colaborativas actuales, las unidades de negocios y equipos pueden trabajar juntos en cualquier momento, ya sea desde oficinas contiguas o en cualquier parte del mundo. Estas tecnologías hacen posible que se puedan contemplar objetivos operacionales como la disminución de tiempos, la racionalización o coordinación de procesos, el recorte de costos y mejoras en el tiempo de comercialización. No obstante, en el uso de herramientas colaborativas, se debe tener presente la gestión de registros, captura de conocimiento y requisitos de

[1] http://en.wikipedia.org/wiki/Sarbanes%E2%80%93Oxley_Act
[2] http://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act
[3] http://en.wikipedia.org/wiki/Design_Criteria_Standard_for_Electronic_Records_Management_Software_Applications

cumplimiento de normas (por ejemplo, para una industria todas las comunicaciones con el cliente deben ser guardadas).

Costos: aunque es difícil identificar el retorno de inversión directamente, no es imposible ver el impacto de un proceso mejorado de negocios. Las herramientas ECM pueden hacer que las organizaciones sean más eficientes y reduzcan sus costos, al organizar la información para su posterior recuperación, uso y por último descarte.

Continuidad: a menudo mencionado como recuperación de desastres (*disasterrecovery*), el planeamiento de continuidad de negocios es una estrategia para asegurar que las operaciones prosigan luego de una interrupción natural o provocada por el hombre. Teniendo en cuenta que los documentos electrónicos son el alma de la mayoría de los negocios, ECM juega un rol clave en la continuidad. Las tecnologías ECM permiten la creación de repositorios centralizados donde reside toda la información vital para la organización.

Cuando se utilizan metodologías tradicionales, en las que los sistemas son desarrollados ad-hoc para resolver cuestiones puntuales y/o brindar servicios que aporten valor agregado sobre información existente, se observa que cada desarrollo es llevado a cabo de manera independiente y luego resulta difícil, costoso o casi imposible integrar los distintos sistemas desarrollados. Así mismo, es complicado y laborioso, atacar los distintos aspectos que posee la información: publicación en la web, mantenimiento de registros históricos, auditoría, etc.

Surge la necesidad y conveniencia de utilizar otras soluciones en las que la captura de información, su clasificación, registro, flujo, y disposición final sea brindada como un servicio sobre el que se pueden construir todas las aplicaciones. [1]

En este trabajo nos propusimos estudiar el campo de los sistemas de Gestión de Contenido Organizacional y su aplicación en el desarrollo de sistemas.

Los beneficios que brindan estos sistemas y la experiencia de haber utilizado metodologías tradicionales, en la que los sistemas son desarrollados ad-hoc para resolver cuestiones puntuales y/o brindar servicios de valor agregado sobre información existente, y en la que cada desarrollo es llevado de manera independiente resultando su integración muy difícil y costosa, han sido el principal factor de motivación.

2 Necesidad y Oportunidad

Hay numerosos factores que llevan a las organizaciones a adoptar las soluciones ECM, por ejemplo el aumento de la eficiencia, la mejora en el control de la información y la reducción de los costos generales en su gestión. Las aplicaciones ECM facilitan el acceso a los registros a través de palabras claves y búsquedas de texto completo, brindando así a los usuarios la información necesaria directamente desde sus puestos de trabajo en segundos en lugar de acceder a múltiples aplicaciones o búsqueda manual por medio de registros en papel.

Los sistemas ECM pueden reducir las necesidades de almacenamiento, papel y envío de correspondencia, haciendo más eficientes el trabajo de sus usuarios y resultando en mejores decisiones para la organización, lo que deriva en una reducción de costos del manejo de la información.

Al comienzo hablamos de las estrategias, métodos y herramientas para ECM. Particularmente, nos concentraremos en un software gestor de contenido organizacional.

Un gestor de contenidos organizacional es una herramienta que permite la gestión de grandes cantidades de información almacenadas en forma de documentos.

La combinación de este tipo de bibliotecas de documentos con índices almacenados en una base de datos permite el acceso rápido a la información mediante diversos métodos. Dicha información está contenida en los documentos que generalmente se encuentran comprimidos y que, además de texto, pueden contener cualquier otro tipo de documentos multimedia (imágenes, videos, etc.).

Un gestor de contenidos organizacional posibilita compartir la información contenida en los documentos que son creados, editados y borrados por sus usuarios. Por consiguiente debe proveer de mecanismos que posibiliten esta colaboración y todo lo que ello conlleva: organización del repositorio de documentos, gestión de los usuarios y sus permisos para el acceso a los documentos y para la modificación de estos, control de versiones de documentos, búsquedas, etc. [2]

El gestor de contenidos organizacional, además de estas características básicas puede ofrecer otras como: notificaciones a los usuarios, reglas de publicación de documentos, mecanismos avanzados de creación de documentos a partir de plantillas, etc.

Actualmente las tecnologías que componen un ECM son descendientes de los sistemas EDMS [4](*ElectronicDocument Management Systems*) de fines de los 80 y principios de los 90. Estos eran productos autónomos (*standalone*) que proveían funcionalidad en una de cuatro áreas: captura de imágenes (*imaging*), flujos de trabajo (*workflow*), gestión documental (*documentmanagement*) y registros en discos ópticos (COLD/ERM).

La primera implementación típica incluía un sistema de captura de imágenes y flujos de trabajo a pequeña escala, posiblemente para un único departamento, con el objetivo de mejorar un proceso que requería mucho papeleo (*paper-intensive*) y el traslado hacia la mítica "oficina sin papel".

Estos primeros sistemas autónomos fueron diseñados para ahorrar tiempo y/o mejorar el acceso a la información reduciendo las necesidades de manejo y almacenamiento de papel, y por consiguiente también se vio reducida la pérdida de documentos y se facilitó el acceso online a la información, que previamente solo estaba disponible en papel, microfilm o microfichas. Las trazas de auditoría de estos sistemas mejoraron la seguridad de los documentos y proveyeron métricas que ayudaron a medir la productividad e identificar la eficiencia.

A fines de los 90, estas tecnologías captaron la atención a las organizaciones que necesitaban soluciones enfocadas y tácticas para tratar problemas claramente definidos.

Con el transcurrir del tiempo, las organizaciones obtuvieron áreas de productividad con estas tecnologías, y quedó claro que las distintas categorías de productos EDMS eran complementarias. Debido a ello, se consideró necesaria y conveniente una amplia adopción de estos productos en toda la organización. La industria ofrecía múltiples soluciones autónomas EDMS, con poca o ninguna integración. Pero a

[4]https://en.wikipedia.org/wiki/Edms

finales de los 90, la integración se incrementó. A partir del año 2001, la industria comenzó a utilizar el término *Enterprise Content Management* para referirse a aquellas soluciones integradas.

En el año 2006, Microsoft con su producto Sharepoint y Oracle Corporation con Oracle Content Management se unieron a líderes establecidos de la industria tales como EMC Documentum, Laserfiche y Open Text, para ofrecer productos en el segmento *entry-level* de ECM.

Existen además productos open source, tales como Alfresco, LogicalDoc, Sense/net, eZpublish, KnowledgeTree, Jumper 2.0, Nuxeo y Plone.

En la actualidad, las organizaciones pueden desplegar un solo sistema ECM que flexiblemente gestiona la información de todos sus departamentos. Combinando ECM con soluciones de BI[5], generan sistemas EIM (enterpriseinformationmanagement), para la gestión de la información empresarial, tanto estructurada como no estructurada.[3]

3 Análisis y Comparación

3.1 Alfresco

Es un sistema de gestión de contenido organizacional gratuito, para sistemas operativos MS Windows y Unix-like. Se ofrece en tres versiones:

CommunityEdition: como software gratuito con licencia LGPL, de código abierto y que cumple con estándares abiertos.

EntepriseEdition: como software propietario y comercialmente licenciado de código abierto y que cumple con estándares abiertos.

In the Cloud: es la versión SaaS[6] de Alfresco.

Incluye un repositorio de contenido, un framework de portal web para gestionar y utilizar contenido estándar de portal, una interfaz CIFS[7] que provee compatibilidad de sistema de archivos con sistemas operativos Windows y Unix-like, un sistema de gestión de contenido web capaz de virtualizar aplicaciones web y contenido estático vía Apache Tomcat, indexación de texto completo de contenidos vía Lucene, y flujos de trabajo con el motor Activiti.

En el núcleo de un sistema Alfresco hay un repositorio que tiene el soporte de un servidor que almacena contenido, metadatos, asociaciones e índices de texto completo. Las interfaces de programación soportan múltiples lenguajes y protocolos con los que los desarrolladores pueden crear aplicaciones personalizadas y soluciones. Las aplicaciones *out-of-the-box* proveen soluciones de gestión documental, colaboración, gestión de registros y archivo, y gestión de contenido web.

Alfresco es en la actualidad la compañía más grande de gestión de contenido open source del mundo, con más de 7 millones de usuarios que utilizan versiones Enterprise, Cloud, Mobile y Community para gestionar 4.000 millones de documentos, y se utiliza en 182 países. [4][5]

[5]Business Intelligence

[6] Software as a Service

[7] Common Internet File System

3.2 Nuxeo

NuxeoEnteprisePlataform (Nuxeo EP) es una plataforma de sistemas de gestión de contenidos de código abierto usados por arquitectos y desarrolladores para construir, desplegar y ejecutar aplicaciones de negocios centradas en contenido. Puede ejecutarse en ambientes Windows y Unix-like.

Provee una plataforma de código abierto basada en java que es modular y extensible para el desarrollo de software ECM y que incluye un conjunto de módulos empaquetados para gestión documental, colaboración, gestión de activos digitales y gestión de casos. Entre otros, provee estos beneficios:

- Extensibilidad: arquitectura altamente flexible basada en sistemas de componentes y orientadas a servicios
- Escalabilidad: tiene buen escalamiento adaptándose a las necesidades del proyecto.
- Adaptabilidad: con Nuxeo Studio, un ambiente de personalización y configuración
- Desarrollo rápido de aplicaciones (*RAD*): manejado por una plataforma modular con componentes reutilizables.
- Código abierto, estándares abiertos: significa un compromiso a la interoperabilidad y la innovación manejado por la comunidad. [6]

3.3 SharePoint

Microsoft SharePoint es una plataforma de aplicaciones Web desarrollada por Microsoft. Lanzado por primera vez en 2001, SharePoint se ha asociado históricamente con la gestión de contenidos de la Intranet y gestión de documentos, pero con el transcurso del tiempo hoy ofrece capacidades mucho más amplias.

SharePoint está construido sobre Microsoft .Net, ASP.Net, Internet Information Server y SQL Server. Todas las plataformas deben correr en un servidor con Windows 2008 o superior de 64 bits.

Existen tres versiones:

- SharePoint Foundation
- SharePoint Standard
- SharePoint Enterprise [7]

4 Comparación de Características

Luego de analizadas en forma no exhaustiva las características de los software ECM, es posible realizar una comparativa desde el punto de vista de su forma de uso y para qué fueron concebidos.

Las alternativas open source a priori parecen no tener costos asociados a su despliegue. Sin embargo, requieren personal de IT altamente formado o bien la contratación de servicios de consultoría y soporte para lograr una implementación exitosa.

SharePoint: es un producto completo que dependiendo del tipo de licencia puede ser gratuito (Foundation) o pago, para llegar a las características empresariales (Server Enterprise) con incorporación de herramientas de BI y un ambiente totalmente integrado en el universo Microsoft. Como punto a favor: si la organización ya ha

invertido en tecnologías Microsoft, incorporar SharePoint en cualquiera de sus versiones es una transición de bajo impacto en el ambiente TIC y en los usuarios. Usa otros productos Microsoft como infraestructura (Windows Server, SQL Server, IIS) por lo que todo el conocimiento de estas plataformas es aplicable para el despliegue de SharePoint. Se ofrece un completo ambiente de desarrollo (VS 2010) en el que las aplicaciones SharePoint son reconocidas y desplegadas fácilmente.Como aspectos negativos podemos citar:

- Si bien existe una versión gratuita su funcionalidad es limitada y características como por ejemplo el modelado de contenido, vistas previas de resultado de búsquedas y tableros de control del usuario solo están disponibles en la versión Enterprise.[8]
- Su principal fortaleza es también una debilidad ya que si no se ha invertido en software de back-end Microsoft el despliegue de cualquiera de las versiones de SharePoint requerirá una inversión en los productos de infraestructura y también una atadura a un proveedor de software en particular (vendorlock in).
- Cada cliente que accede a un servidor SharePoint debe tener una licencia de acceso (CAL[8]) por lo que hay que sumarlo al costo total (TCO[9]).

Nuxeo: en pocas palabras, es una plataforma de desarrollo más que un producto terminado. Su disponibilidad de software open source con todas las características hace viable una evaluación completa. Permite independizarse del ambiente de ejecución y no tiene ataduras con ningún tipo de software de infraestructura. Su modelo de despliegue es adecuado cuando desea desarrollarse aplicaciones de ECM verticales. En el modelo de suscripción pago no se pagan licencias de acceso de clientes, y se brinda soporte técnico, mantenimiento de software y herramientas de personalización.Como aspectos negativos podemos citar:

- Su modelo de licenciamiento obliga a pagar por las herramientas de desarrollo o de otra forma hay que recurrir a la edición de archivos XML para configurar distintos aspectos de la aplicación: branding, workflows, interfaces de usuario, etc.
- La comunidad de usuarios y desarrolladores no es tan extendida como en el caso de Alfresco.
- No tiene herramientas de BIout-of-the-box.

Alfresco: al igual que Nuxeo, al ser software open source permite una evaluación completa de sus características. Posee una interfaz web de usuario rica y simple, con mucha potencia y capacidades de extensión a través de *gadgets* que se incorporan en los tableros de control del usuario. También permite independizarse del ambiente de ejecución. La comunidad Alfresco es muy amplia y activa brindando un gran número de *addons* gratuitos o pagos. Es una aplicación madura con una gran base de usuarios, y es el software ECM open source más utilizado en el mundo. Su modo de licenciamiento *on- premise* no requiere de licencias de acceso al cliente y ofrece soporte técnico y mantenimiento.Como aspectos negativos podemos citar:

- No tiene herramientas de BI out-of-the-box.
- No posee herramientas de desarrollo integradas gratuitas. Existen addons con funcionalidad limitada y hay que recurrir a la edición de archivos XML para otras configuraciones.

[8] Client Access License

[9] Total Cost of Ownership

En la tabla 1 se resumen las características que a nuestro criterio pueden resultar de interés a la hora de la toma de decisiones sobre la elección de una herramienta sobre otra para un proyecto en particular.

Tabla 1. Comparación de características

Característica	Alfresco	Nuxeo	SharePoint
Open source	Si	Si	No
Versión gratuita	Si	Si	Si[10]
Vendorlockin	No	No	Si
Soporte de BI integrado	No	No	Si
Licenciamiento por cantidad de clientes	No	No	Si
Gestión del ciclo de vida de la información	Si	Si	Si
Cumplimiento de regulaciones de gestión de registros (HIPAA, Sarbanes-Oxley, e-discovery)	Si	No	Si
Versión *in thecloud*	Si	Si	Si
Herramientas de desarrollo gratuitas	Si[11]	No	Si
Herramientas de desarrollo provistas por el fabricante	No	Si	Si
Herramientas de administración completas e integradas (backup/restore, gestión de componentes, aplicaciones, etc)	No	No	Si
Trabajar desconectado (cliente de sincronización offline)	Si	Si	Si

5 Desarrollo sobre Alfresco

Procedimos a desarrollar una aplicación pequeña sobre Alfresco, con el objetivo de evaluar el nivel de dificultad que esto conlleva. Elegimos un organismo no informatizado dentro de nuestro ambiente laboral del Poder Judicial como escenario para la implementación de un sistema informático. El organismo elegido es la Dirección de Mediación del Poder Judicial de la Provincia del Chubut, que es la encargada de organizar el Registro Provincial de Mediadores y el Servicio Público de Mediación. Por razones de espacio, en este trabajo no se describe la problemática con mayor detalle, sólo se quiere mencionar el resultado de la experiencia.

Habiendo definido el modelo de datos de acuerdo al dominio de aplicación, observamos que rápidamente podemos dar soporte a múltiples requerimientos con las capacidades out-of-the box: subir contenido al repositorio, realizar búsquedas personalizadas, streaming de contenido, gestionar usuarios, permisos de acceso, diseñar micrositios con paneles personalizados, acceder y actualizar contenido mediante protocolo CIFS, asignar y modificar metadatos al contenido, etc.

[10] Con funcionalidad limitada y requiere una licencia de Windows Server.
[11] Addons de terceras partes, con capacidades limitadas

Para la implementación de la aplicación mencionada anteriormente elegimos herramientas gratuitas y open source. La dificultad que esto conlleva es la obtener un ambiente productivo con rapidez, teniendo que ensamblarse parte por parte. También es dificultosa la obtención de documentación de algunas herramientas, y sólo se dispone de los foros de usuarios. Sin embargo, la experiencia es altamente favorable y nos alienta a utilizar el modelo de gestión de contenido para el desarrollo para futuras aplicaciones.

6 Conclusiones

Uno de los problemas que todas las organizaciones enfrentan en la incorporación de TIC es la dispersión de los contenidos digitales y la gestión del ciclo de vida desde su creación o captura hasta la disposición final. También deben tenerse en cuenta cuestiones como la correcta catalogación, almacenamiento, preservación, distribución, su búsqueda y el seguimiento de políticas de acceso a la información. Esto representa un desafío en cada nuevo sistema a desarrollar. Así mismo, estos sistemas deben integrarse y colaborar con otros existentes.

Es por ello y tal como indicamos al inicio del presente trabajo que la gestión del contenido organizacional es una necesidad, constituida no sólo por herramientas de gestión de contenido, sino también por estrategias y métodos. Así podremos hacer frente a la gestión de la información no estructurada, donde quiera que la información se encuentre.

Habiendo analizado tres productos testigo de gestión ECM tanto en el campo comercial como open source hemos podido apreciar sus potencialidades para ser utilizados como infraestructura de base en la creación de nuevos sistemas.

Las herramientas analizadas cumplen con una gran gama de requerimientos funcionales que se presentan en la gestión de contenido organizacional. Out-of-the box proveen las siguientes características, entre otras:

- Gestión del ciclo de vida de la información
- Cumplimiento de normas de gestión de registros
- Acceso desde dispositivos móviles
- Sincronización con clientes desconectados
- Selección de implementación on-premise/in thecloud
- Cumplimiento de políticas organizacionales de acceso a la información

La elección del producto dependerá de:

1. la infraestructura de software back-end existente
2. el recurso humano disponible de IT, su formación y experiencia
3. el recurso financiero asignado a la implementación del proyecto

Para el caso mostrado, la elección recayó en Alfresco basados en un conocimiento previo del software, el deseo de utilizar herramientas open source y la idea general que Alfresco es un producto más maduro y completo que la otra alternativa open source (Nuxeo). Sin embargo, cualquiera de ellos se adecuaba al propósito planteado.

Con la experiencia de haber utilizado con anterioridad metodologías tradicionales en las que los sistemas son desarrollados ad-hoc, reconocemos la importancia de este tipo de herramientas que brindan soluciones robustas para la gestión de información,su clasificación, registro, flujo, y disposición final. Que esto sea brindado

como un servicio sobre el que sepueden construir todas las aplicaciones resulta ser una solución muy superadora.

7 Referencias

[1] What is ECM? What is Enterprise Content Management?, http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx
[2] Manual de Usuario de Alfresco, https://documenta.ugr.es:8443/alfresco/d/d/workspace/SpacesStore/b155c802-f825-4e49-9ab6-256625c7d2be/ManualAlfresco_CSIRC_v1.0.pdf
[3] Enterprise Content Management. http://en.wikipedia.org/wiki/Enterprise_Content_Management
[4] Alfresco (software), http://en.wikipedia.org/wiki/Alfresco_(software)
[5] Alfresco Accelerates Growth and Adds Over 500 Enterprise Customers, http://www.alfresco.com/news/press-releases/alfresco-accelerates-growth-and-adds-over-500-enterprise-customers
[6] Enterprise Content Management Platform for Business Apps, http://www.nuxeo.com/en/products/content-management-platform
[7] Microsoft SharePoint, http://en.wikipedia.org/wiki/Microsoft_SharePoint
[8] Khamis, N. SharePoint 2010 Standard vs Enterprise vs Foundation edition features . http://www.khamis.net/blog/Lists/Posts/Post.aspx?ID=61

Marcos Metodológicos dentro de la Informática

Pablo Iuliano[1], Luis Marrone[1], and Elvio Fernandez

LINTI, Facultad de Informática UNLP,
La Plata, Argentina
{piuliano,lmarrone}@info.unlp.edu.ar

Abstract. El presente documento abordará e indentificará dos corrientes filosóficas contrapuestas dentro de la ciencia, el Racionalismo y el Empirísmo, en relación con la actividad informática. En primera instancia se describirán las corrientes filosóficas antes citadas, para luego realizar un análisis en cual se vinculará las corrientes filosóficas a casos concretos dentro de la disciplina en cuestión. Finalmente se presentarán las conclusiones obtenidas.

Keywords: Racionalismo; Empírismo; Informática; Métodos;

1 Marco histrico-filosfico

En este apartado se introducirán las tendencias filosóficas de interés para este trabajo. Lejos de ser un estudio comparativo de distintas ramas de la filosofía de la ciencia se tocarán los aspectos más representativos y necesarios para su posterior vinculación con la visión y la actividad de la informática.

1.1 El siglo XVII

Paralelamente con las leyes de Johannes Kepler sobre el movimiento de los planetas y sus órbitas y el descubrimiento por Galileo Galilei de la ley de caída libre de los cuerpos se desarrollan los escritos de Francis Bacon, en los que enfatiza su idea de *gran renovación* de la ciencia a partir de la definición de un *nuevo instrumento* que aparta los prejuicios de los espíritus de los hombres. Esta crítica de Bacon, a la vieja ciencia, es la precursora de dos grandes vertientes de la *nueva filosofía* del siglo XVII. El racionalismo cartesiano y la interpretación empirista de la naturaleza.

1.2 Racionalismo

"En el siglo XVII la comunidad de posturas, en lo que respecta a la posibilidad del conocimiento y a su esencia, no se da en la cuestión del origen del conocimiento. El racionalismo coloca a la razón en su origen al aseverar que el pensamiento es la fuente y el fundamento del conocimiento. Ello es así porque responde a las exigencias de necesidad lógica y de validez universal..." [1].

Para argumentar esta *validez universal* el racionalismo plantea que la razón proporciona juicios que pueden ser demostrables. Aquellas expresiones de las que se puede decir que son evidentemente y necesariamente ciertas o validas o, por el contrario, que son forzosamente falsas o invalidas, las cuales encierran una verdad necesaria. Mientras que las experiencias (emociones y sensaciones) proporcionan juicios que expresan únicamente la posibilidad/ambigüedad de algo, y de las cuales pueden pensarse lo contrario. De esta manera el racionalismo enuncia que el verdadero conocimiento y la naturaleza de la realidad tiene un único y necesario origen: el racional.
El innatismo postulado por Descartes, enuncia que el entendimiento es fundado en una ilustración del espíritu por Dios. Existen ideas innatas a partir de las cuales los hombres establecen relaciones lógicas entre proposiciones que constituyen parte de la actividad de la razón.
"... He notado ciertas leyes que Dios ha establecido en la naturaleza y cuyas nociones ha impreso en nuestras almas, de tal suerte que, si reflexionamos sobre ellas. Con bastante detenimiento, no podremos dudar de que se cumplen exactamente en todo lo que es o se hace en el mundo" [2].
La *idea innata* es solo una de las formas de lo que Descartes llama sustancia. La sustancia existe de tal manera que no tiene necesidad de otra cosa para existir [2]. Se afirma la existencia de tres sustancias:

1. La sustancia finita pensante (el ser pensante).
2. La sustancia extensa (mundo)
3. El propio Dios, sustancia infinita pensante.

Son tres sustancias pero sólo dos modos de ser sustancia: el pensamiento y la extensión o materia. Resulta relevante el término de *sustancia* si lo traducimos con el concepto de idea. Especialmente la noción de la existencia de la idea de *"yo como sujeto pensante"*. A diferencia de la sustancia extensa, como la idea del mundo como objeto del conocimiento; en cuanto a Dios, es garantía que nuestro conocimiento está en perfecta concordancia con el orden de la realidad. Es Dios quien pone en forma de ideas innatas los pensamientos claves para que el hombre alcance el conocimiento a través de la razón.
En la física y las demás ciencias basadas en el estudio de la realidad extensa, lo mensurable y analizable según conceptos geométricos, Descartes aplico el modelo teórico aportando la hipótesis de que el mundo inanimado y los cuerpos animados son máquinas (mecanicismo), y que todos los fenómenos físicos son explicables fundamentados en este principio. El fuerte carácter de la filosofía mecanicista cartesiana fascinó a los estudiosos de la época, incluso a los empiristas [1].
La influencia del mecanicismo continúa hasta la primera revolución industrial -situado desde el siglo XVII al XIX- teniendo como símbolo los engranajes [3]. Puramente mecanicista, la organización es vista como una máquina solamente para satisfacer los deseos de lucro. Sólo posteriormente en la segunda revolución se puede observar el desarrollo de los procesos industriales devenidos de *engranajes* a *cadena de montajes*.

1.3 Empirismo

En contraposición, con el pensamiento cartesiano, el empirísmo rechaza (antiinnatismo) de raíz todo iluminismo, de toda fuerza trascendente que pueda postularse como ilustrativa del alma humana.
Para Thomas Hobbes no existe el dualismo mente-cuerpo que enuncia el pensamiento cartesiano. Esta dualidad es resuelta mediante la reducción de ambos en un único principio material. Las características propias de la mente no escapan a las leyes que rigen las demás cosas. Entonces, la fuente para conocer los objetos materiales está en las sensaciones suministradas por nuestros sentidos, los cuales constituyen el único criterio de verdad. Así, la razón opera únicamente con nombres, que no son otra cosa que pura convención, y el lenguaje es el instrumento necesario para la adquisición del conocimiento. Ello sucede recordando los hechos de la experiencia, operando sobre ellos según ciertas reglas y comunicándolos a otros individuos. Razón y lenguaje, desde este punto de vista, aparecen tan artificiales o convencionales como lo pueda ser la sociedad.
La no admisión de conceptos universales, ni en las cosas ni en la mente, conduce a la doctrina sensualista de Hobbes a definir toda actividad intelectiva como un puro movimiento. Entendiendo este, como la actividad más primordial/básica en la mente, aquello que dispara la imaginación siendo pensado como un movimiento en los órganos del cuerpo.
John Locke, añade que las ideas simples proceden directamente de la experiencia, para alimentar así la *tabla rasa* de la mente. Y las ideas complejas se componen de las anteriores mediante su combinación, yuxtaposición o abstracción. Estas operaciones mentales se llevan a cabo según leyes aportadas por la experiencia; la razón es guía de todo conocimiento probable y no tiene otro límite que la experiencia. De esta manera Locke señala que el conocimiento posible solo puede suceder a *posteriori*. El conocimiento es basado en la experiencia.
A diferencia del empirismo, el racionalismo no ve limite en la razón. Ésta, con los métodos adecuados, permitiría el abordaje y entendimiento del *todo*. El empirismo, por el contrario, sitúa al hombre como alguien acotado forzando un corrimiento en el foco desde el hombre hacia el fenomeno/contexto a observar. La experiencia es el límite en el entendimiento de la complejidad de la realidad.

2 Racionalismo dentro de la Informática

Como se verá a continuación, dentro de la informática aún son notables las influencias de las ideas y abordajes del siglo XVII. El mecanicismo y racionalismo de Descartes están entrelazados de manera indivisible en la historia de la informática y por consiguiente en los métodos y técnicas del quehacer informático.

2.1 Desarrollo en Cascada

Una de las técnicas más utilizadas a lo largo de la historia del desarrollo de software es la conocida como desarrollo en cascada o modelo en cascada. Este

es un enfoque metodológico que ordena rigurosamente las etapas del proceso de desarrollo de software, de tal forma que el inicio de cada etapa debe esperar a la finalización de la etapa anterior. El desarrollo en cascada esta compuesto por las siguientes fases o etapas:

1. Análisis de requisitos: En esta etapa se analizan los requerimientos que deberá cumplir el software mediante una o varias entrevistas con el o los usuarios a los cuales esta dirigido el software y así determinar qué objetivos debe cumplir el mismo. De esta fase surge un documento con la especificación completa de lo que debe hacer el sistema sin entrar en detalles de implementación. Cabe señalar que en esta etapa se consensua todo lo que se requiere del sistema y no pudiéndose solicitar nuevos requisitos a mitad del proceso de elaboración del software.
2. Diseño del sistema: Aqui se descompone y organiza el sistema en elementos que puedan elaborarse por separado, aprovechando las ventajas del desarrollo en equipo. Como resultado surge un documento, que contiene la descripción de la estructura relacional global del sistema y la especificación de lo que debe hacer cada una de sus partes, así como la manera en que se combinan unas con otras.
3. Codificación: Es la fase en donde se implementa el código fuente, haciendo uso de un lenguaje de programación. Dependiendo del lenguaje utilizado y su versión se crean las bibliotecas y componentes reutilizables dentro del mismo proyecto para hacer que la programación sea un proceso mucho más rápido.
4. Pruebas y Verificación: Los elementos, ya programados, se ensamblan para componer el sistema, se comprueba que funciona correctamente y que cumple con los requisitos, antes de ser entregado al usuario final.
5. Mantenimiento: Es la etapa más extensa de todo el proceso de desarrollo. El sistema se instala y se pone en funcionamiento corrigiendo todos los errores no descubiertos en las etapas anteriores y se realizan mejoras de implementación. Llegado el caso se identifican nuevos requisitos.

Reminiscencias racionalistas en el desarrollo en cascada En proyectos de desarrollo de software de la vida real, rara vez se sigue una secuencia lineal, esto crea un deficiente analisis de requisitos lo cual es el punta pie inicial para obtener una mala implementación del sistema y concluyendo en una etapa de mantenimiento practicamente insostenible, lo cual generalmente conduce al fracaso.
Al ser un proceso lineal-secuencial que no permite la reformulación de etapas ya transitadas, lo diseñado en las etapas tempranas del proyecto es lo que será finalmente el sistema consumado.
Existe una sola dirección en la formulación del análisis/diseño. No existe ninguna etapa de contrastación con la realidad, asumiendo la actividad del desarrollo como un proceso inyectivo análisis/diseño $\Rightarrow$ codificación/mantenimiento.
Puede notarse en este tipo de abordaje las reminiscencias del pensamiento racionalista donde la razón era la única herramienta valida para las construcciones int-

electuales. Este modelo de desarrollo de software posiciona las etapas asociadas al raciocinio del sistema en las etapas más tempranas del proceso y relegando la acción de construcción para el final del mismo. Así se deja entre ver que el desarrollo en cascada esta inspirado en la frase *"Pienso luego Existo"* o en este caso *"Pienso luego desarrollo"*. Otra consideración relevante es que la etapa del mantenimiento se posiciona en el final del proceso de desarrollo, la cual termina siendo una etapa de adecuación de lo ideado a la realidad.

3 Empírismo dentro de la Informática

Las aproximaciones empiristas dentro de la informática muchas veces están relacionadas con los abordajes de la complejidad y los sistemas abiertos [4]. Lo estudiado no es fácilmente asimilable debido a la naturaleza no lineal del objeto de estudio.
Esta última afirmación queda debidamente evidenciada a partir de los casos presentados a continuación.

3.1 Desarrollo de Aplicaciones de Simulación

Las aplicaciones orientadas a intentar resolver problemas complejos, o al menos proveer un medio para poder entenderlos, son en general implementadas como herramientas computacionales de simulación.
Pero antes de describir las posibles opciones para implementar herramientas de este tipo se debe definir que se entiende por simulación, para dicho fin a continuación se citará una definición perteneciente a Shannon Robert para el concepto antes mencionado:
"La simulación es el proceso de diseñar y desarrollar un modelo computarizado de un sistema o proceso y conducir experimentos con este modelo con el propósito de entender el comportamiento del sistema o evaluar varias estrategias con las cuales se puede operar el sistema"[5].
De la definición anterior surgen dos conceptos, modelo y proceso:

1. Modelo de simulación: conjunto de hipótesis acerca del funcionamiento del sistema expresado como relaciones matemáticas y/o lógicas entre los elementos del sistema.
2. Proceso de simulación: ejecución del modelo a través del tiempo en una computadora para generar muestras representativas del comportamiento.

La simulación como método empírico Claramente las aplicaciones de simulación tienen una naturaleza netamente empírica ya que no se conoce de antemano los resultados a los cuales se podrán arribar antes de su ejecución. Así como la construcción de un simulador también implica transitar un proceso sistematico de prueba y error que se reitera tantas veces hasta que la aplicación se ajuste lo suficientemente al problema del mundo real que se intenta responder o entender según sea el caso.

3.2 Transición de IP versión 4 a IP versión 6

Las mayoria de las redes pequeñas, medianas y grandes funcionan utilizando un protocolo de red llamado IP (*Internet Protocol*). Este protocolo de red es en el cual se basa la transmisión de datos en Internet y su definición se encuentra especificada en la RFC 971.[7]
Para que los equipos, por ejemplo computadoras e impresoras, que forman parte de una red IP se puedan comunicar, cada uno de ellos debe tener asignado una dirección IP que los identifique univocamente y así poder establecer comunicación con otros equipos.
La versión que se usa actualmente para el direccionamiento IP es IPv4. Esta versión cuenta con direcciones de 32 bits de longitud, con lo cual se obtiene un total de 4.294.967.296 direcciones disponibles. Dentro de las cuales, hay un gran número de ellas destinadas para usos reservados y no se pueden utilizar para ser asignadas a dispositivos de red.

El problema con IPv4 El surgimiento de nuevas tecnologías y de nuevos dispositivos que requieren conectividad a Internet, han hecho que la demanda de direcciones IP crezca de forma grotesca. Smartphones, camaras con acceso a internet, etc. han hecho que la vida útil de esta versión se reduzca de forma drástica cada vez más. En un principio se pronosticó que las direcciones IP se acabarían cerca del año 1995, pero gracias a diversas técnicas se ha pospuesto dicha proyección para finales del año 2013 aproximadamente.

La solución al agotamiento de direcciones IP Teniendo en mente la problematica del agotamiento de las direcciones IP y después de examinar diferentes propuestas, la Internet Engineering Task Force (IETF) decidó adoptar la propuesta recomendada en 1995 y definida en el RFC 1752.[8] Esta propuesta aparece como una nueva versión del protocolo IP y debido a que se realizaron varias pruebas, extensiones y modificaciones a la versión 4 con el fin de resolver el problema en cuestión, para evitar confusiones futuras a la nueva versión se le asignó el número 6. De esta manera surgió la próxima generación de IP denominada IPv6. La caracteristica más relevante de IPv6 es que provee direcciones de 128 bits de longitud y con esto desaparece por completo el problema de agotamiento de direcciones ya que por ejemplo existirán 295 direcciones por cada ser humano existente en el planeta.

La transición de IPv4 a IPv6 Para la mayoría de los usuarios su necesidad de conectividad radica en el acceso a la web, correo electrónico y a su conjunto de aplicaciones particulares (Facebook, twitter,etc.) y esta necesidad debe estar cubierta los 7 días de la semana las 24hs del día. Por lo tanto para realizar un buen despliegue de IPv6 en cualquier infraestructura de red, nos enfrentamos al reto de lograr la misma de manera que transparente para el usuario (o sea que no experimente ninguna suspensión en la conectividad a la red) y no como una

experiencia traumática. Es por ello que toda transición o migración de tecnología debe estar planificada sobre los siguientes tres pilares fundamentales:

1. Planificar el despliegue de manera precisa por fases.
2. Realizar pruebas sobre el diseño de la nueva infraestructura para evaluar el funcionamiento de éste y detectar cualquier problema antes de desplegar la nueva red.
3. Desplegar la red en el entorno de producción.

Una vez alcanzada la finalización del proceso de migración de la red a IPv6, seguramente estaremos en presencia de un escenario heterogéneo donde las redes con las cuales establecia conexión la red migrada siguen funcionando con la versión anterior. Es por ello que existen varias técnicas que posibilitan la coexistencia entre redes con distintas versiones de IP y asi lograr el aseguramiento del servicio a los usuarios.

La transición tecnológica como ciclo contrastante de la realidad En todo proceso de migración o transición de tecnología no se puede asegurar a priori que las fases ideadas no se vean a afectadas por fallos inesperados, es por ello que este tipo de proceso se realiza paulatinamente en entornos controlados y en ciclos de puesta en producción, testeo, mantenimiento y vuelta a producción. Estos ciclos contrastantes van generando conocimiento a posteriori de cada iteración como es propuesto por el empirísmo.

3.3 Aplicaciones Inspiradas en la Biología

Muchas de las líneas de investigación en la informática se han inspirado en la biología, en esta sección presentaremos las dos más relevantes y como estas se relacionan con el empírismo.

Algoritmos Genéticos *Un algoritmo genético (AG) es una técnica de programación que imita a la evolución biologíca como estrategia para resolver problemas. Dado un problema específico a resolver, la entrada del AG es un conjunto de soluciones potenciales a ese problema, codificadas de alguna manera, y una métrica llamada función de aptitud que permite evaluar cuantitativamente a cada candidata. Estas candidatas pueden ser soluciones que ya se sabe que funcionan, con el objetivo de que el AG las mejore, pero se suelen generar aleatoriamente. Luego el AG evalúa cada candidata de acuerdo con la función de aptitud. En un acervo de candidatas generadas aleatoriamente, por supuesto, la mayoría no funcionarán en absoluto, y serán eliminadas. Sin embargo, por puro azar, unas pocas pueden ser prometedoras -pueden mostrar actividad, aunque sólo sea actividad débil e imperfecta, hacia la solución del problema. Estas candidatas prometedoras se conservan y se les permite reproducirse. Se realizan múltiples copias de ellas, pero las copias no son perfectas; se introducen cambios aleatorios durante*

el proceso de copia. Luego, esta descendencia digital prosigue con la siguiente generación, formando un nuevo acervo de soluciones candidatas, y son sometidas a una ronda de evaluación de aptitud. Las candidatas que han empeorado o no han mejorado con los cambios en su código son eliminadas de nuevo; pero, de nuevo, por puro azar, las variaciones aleatorias introducidas en la población pueden haber mejorado a algunos individuos, convirtiéndolos en mejores soluciones del problema, más completas o más eficientes. De nuevo, se selecionan y copian estos individuos vencedores hacia la siguiente generación con cambios aleatorios, y el proceso se repite. Las expectativas son que la aptitud media de la población se incrementará en cada ronda y, por tanto, repitiendo este proceso cientos o miles de veces, pueden descubrirse soluciones muy buenas del problema.[9]

Redes Neuronales Una red neuronal es un método de resolución de problemas basado en un modelo informático de la manera en que están conectadas las neuronas del cerebro. Una red neuronal consiste en capas de unidades procesadoras, llamadas nodos, unidas por conexiones direccionales: una capa de entrada, una capa de salida y cero o más capas ocultas en medio.
Se le presenta un patrón inicial de entrada a la capa de entrada, y luego los nodos que se estimulan transmiten una señal a los nodos de la siguiente capa a la que están conectados. Si la suma de todas las entradas que entran en una de estas neuronas virtuales es mayor que cierto umbral de activación de la neurona, esa neurona se activa, y transmite su propia señal a las neuronas de la siguiente capa. El patrón de activación, por tanto, se propaga hacia delante hasta que alcanza a la capa de salida, donde es devuelto como solución a la entrada presentada. Al igual que en el sistema nervioso de los organismos biológicos, las redes neuronales aprenden y afinan su rendimiento a lo largo del tiempo, mediante la repetición de rondas en las que se ajustan sus umbrales, hasta que la salida real coincide con la salida deseada para cualquier entrada dada.

Experiencia como fuente del conocimiento Según el empirismo *"la experiencia es la única fuente del conocimiento"*.
Tanto los algoritmos genéticos como las redes neuronales basan su funcionamiento en la adquisición de experiencia. Los primeros a través de los ciclos evolutivos, mientras que los segundos por medio de la fase de entrenamiento. Por lo tanto las redes neuronales como los algoritmos genéticos se encuadran perfectamente como aproximaciones empiristas.

4 Conclusion

Se han examinado casos testigos de acuerdo a la conveniencia y cercanía con las corrientes racionalista y empirista, de manera que sea posible resaltar claramente las influencias en cada caso.

El racionalismo con un matiz netamente antropocéntrico, posiciona al hombre en un lugar central sin límites en el entendimiento del *todo*. El mecanicismo cartesiano reafirma esto y contribuye a la idea consecuente del hombre como creador (de mecanismos) o gran diseñador. Lejos de estar en el ocaso, esta idea ha estado subyacente desde la primera revolución industrial continuando en los refinamientos de los procesos industriales, en las ingenierías y en la joven vida de la profesión informática.

El método de desarrollo en cascada, tal como fue desarrollado en este documento, es una de las primeras formulaciones dentro de lo que en informática se llamó *ingeniería del software*. Este abordaje sufre de precariedad y rigidez, pero tiene asidero en que fue una de las primeras respuestas a lo que en los principios de la década del 70 se llamo *la crisis del software* [11], intentando mejorar los tiempos de creación del software, minimizar los costos, mejorar la verificabilidad y flexibilidad del software desarrollada. Esta primera versión de desarrollo en cascada, adoleció de una gran ingenuidad por parte de los especialistas que generaron el método. Cayendo eventualmente en la misma problemática que se intentaba superar. El método invierte una gran cantidad de energía en definir, pensar y diseñar el andamiaje de lo que después sería el sistema, postergando lo más posible las etapas de constrastación con la realidad. Una vez que el sistema funciona, se pasa a la etapa de mantenimiento en la cual se realiza la adecuación de lo ideado a la realidad.

Cabe destacar que el desarrollo en cascada es usado actualmente pero no en la forma básica y pura desarrollado en este documento. Nuevas versiones, modificaciones y adecuaciones al método de desarrollo en cascada existen. Minimizando así la precariedad y rigidez del primero, aunque no por ello deja de conservar un origen netamente racionalista.

El desarrollo empirista del siglo XVII continua el matiz antropocéntrico del racionalismo, sin embargo asume al hombre como alguien acotado, posicionándolo asi en el lugar de *aprendiz* de la realidad. La realidad es abordable hasta cierto punto y el hombre necesita censar/percibir de alguna manera los fenómenos del mundo. Esto es necesario para empezar la construcción de las implicancias de las actividades sucedidas en realidad, y así, poder formular modelos, ideas y nociones generales de la complejidad del mundo.

Tanto los algoritmos genéticos como las redes neuronales y las aplicaciones de simulación basan su funcionamiento en la adquisición de experiencia. Los primeros a través de los ciclos evolutivos, mientras que los segundos por medio de la fase de entrenamiento. En este tipo de escenarios el desarrollador queda relegado al lugar de observador/supervisor de un proceso autónomo, que se auto regula mediante el censado del contexto. En contraposición los abordajes imperativos quedan excluidos en estos escenarios. No sería factible definir a priori la actividad que *debiera* suceder dentro de este tipo de aplicaciones, ya que el objetivo de éstas es detectar un emergente mediante un gran número de iteraciones y la experiencia obtenida a lo largo del proceso.

En escenarios no-lineales como la implantación de IPv6 los especialistas asumen una óptica más intervencionista, dado que no es del todo claro, ni determinista,

el comportamiento del sistema. En los estadios tempranos de cada intervención es necesario censar la actividad y someter a un análisis exhaustivo los resultados. Posteriormente la intervención en sí, es llevada a cabo de acuerdo a las reformulaciones originadas en el análisis de la actividad percibida. No es difícil observar aquí las influencias netamente empiristas, ni las palabras de John Locke cuando señala que el conocimiento posible solo puede suceder a *posteriori*. Por último es importante destacar hasta qué punto estas corrientes de pensamiento están implicadas en la actividad informática. Aunque en la misma no se mantienen enfoques puristas como los presentados en este documento, el conjunto de actividades informáticas son entonces una yuxtaposición de estos abordajes. El surgimiento de nuevas técnicas para la resolución de problemas dentro de la disciplina, actualmente es una composición de ambas tesituras fijando etapas en las cuales toma un rol más preponderante una para luego cederselo a la otra en una etapa posterior.

References

1. RACIONALISMO Y EMPIRISMO EN LA LINGSTICA DEL SIGLO XVII: JOHN WILKINS Y PORT-ROYAL. Xavier Laborda Gil
2. El discurso del Metodo Rene Descartes. http://www.librosmaravillosos.com/metodo/index.html
3. Analisis de Sistemas- Juan Bravo Carrasco- Cap 4.
4. Complexity and Postmodernism Paul Cilliers Chapter 1
5. Systems simulation: the art and science - Shannon, Robert; Johannes, James D. (1976). IEEE Transactions on Systems, Man and Cybernetics 6(10). pp. 723-724.
6. Connectivity Probability of Wireless Ad Hoc Networks: Definition,Evaluation, Comparison - Tatiana K. Madsen, Frank H. P. Fitzek and Ramjee Prasad
7. A SURVEY OF DATA REPRESENTATION STANDARDS http://tools.ietf.org/html/rfc971
8. The Recommendation for the IP Next Generation Protocol http://tools.ietf.org/html/rfc1752
9. Algoritmos genéticos y computación evolutiva: Adam Marczyk (2004) http://thegeek.org/docs/algen/
10. Red neuronal artificial https://es.wikipedia.org/wiki/Red$_n euronal_a rtificial$
11. The humble programmer by Edsger W.Dijkstra http://www.cs.utexas.edu/ EWD/ewd03xx/EWD340.PDF

Agentes Inteligentes para propiciar la Accesibilidad Web

Gabriela Miranda[1], Adriana Martin[1, 2], Rafaela Mazalu[2, 3]
Gabriela Gaetán[1], Viviana Saldaño[1]

[1] Unidad Académica Caleta Olivia, Universidad Nacional de la Patagonia Austral, Argentina
[2] GIISCo, Facultad de Informática, Universidad Nacional del Comahue, Argentina
[3] Concejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Comahue, Neuquén, Argentina
{gmiranda/ amartin}@uaco.unpa.edu.ar // rafaelamazalu@gmail.com //
{ggaetan/ vivianas}@uaco.unpa.edu.ar

Resumen. A catorce años de las primeras recomendaciones de la W3C, los enfoques que asisten a la accesibilidad de productos Web siguen requiriendo de la intervención y el juicio humano. Si bien existen buenas propuestas que aplican técnicas inteligentes al proceso de evaluación y reparación, la brecha entre el soporte existente y las necesidades reales de automatización para detectar y reparar con mayor precisión las barreras de Accesibilidad, es aún muy significativa. En este trabajo se realiza un profundo relevamiento de los enfoques que dan soporte a la accesibilidad Web exhibiendo rasgos inteligentes. A tal fin, se propone un *Framework de Evaluación* donde se analizan las características de los enfoques para determinar sus fortalezas y debilidades. También y como parte de este trabajo, se propone una *Taxonomía de Agentes* para asistir al proceso de revisión en la identificación y comprensión de los agentes inteligentes.

Palabras Claves: agentes inteligentes, accesibilidad Web, usabilidad universal, recomendaciones de la W3C.

1 Introducción

"*The Visual Web*" parece emerger como una de las características que va a distinguir a la Web 3.0 y enfrenta a nuevos y complejos desafíos para propiciar la accesibilidad. Usuarios cada vez más participativos y con capacidades diferentes y/o necesidades especiales, requieren del despliegue de productos Web que faciliten el acceso a la información, comunicación y servicios.

Por otra parte, el amplio campo de estudio que abarca el desarrollo y aplicación de agentes inteligentes a la Web [13, 21], mejora la interacción hombre-computadora y comparte objetivos con otras áreas de investigación más preocupadas por el perfil humano de la Web, tales como la Usabilidad y la Accesibilidad. En particular, los agentes se incorporan a enfoques de evaluación y reparación de la accesibilidad aplicando estándares internacionales y manipulando diferentes formatos de documentos Web. Sin embargo, aún existe una brecha muy significativa entre el soporte disponible y las necesidades reales en términos de detección y reparación automáticas de las barreras de accesibilidad. Una revisión a la amplia gama de enfoques evidencia que tienen una fuerte dependencia a la intervención y al juicio humano y en general, enfo-

can sus esfuerzos en proveer productos Web accesibles para usuarios con discapacidad visual. En este contexto, se requiere de enfoques de accesibilidad que relacionen y soporten las actividades de evaluación y reparación inteligentemente.

El objetivo de este trabajo es realizar un profundo relevamiento de los enfoques que dan soporte a la accesibilidad exhibiendo rasgos inteligentes para exponer las fortalezas, y enfocar las debilidades donde se deben encausar los nuevos esfuerzos de investigación. En primer lugar, se propone una *Taxonomía de Agentes* para asistir a la identificación y comprensión de los diferentes tipos de agentes inteligentes, la cual puede crecer conforme a las nuevas revisiones del estado del arte. En segundo lugar, se desarrolla un *Framework de Evaluación* basado en el *Análisis de Características* ("*Feature Analysis"*) de la metodología DESMET [6]. Finalmente, se aplica el framework a los 8 (ocho) enfoques seleccionados para determinar el grado de evolución alcanzado por el estado del arte de los enfoques de accesibilidad inteligentes.

Este documento se organiza de la siguiente manera: En la Sección 2 se describe el concepto de "agente inteligente" y se presenta nuestra *Taxonomía de Agentes*; mientras que en la Sección 3 se describen los 8 (ocho) enfoques seleccionados. En la Sección 4 se presenta y aplica el *Framework de Evaluación*. Luego, en la Sección 5 se propone una discusión para determinar el grado de evolución del estado del arte y lo que aún está pendiente de ser explorado y desarrollado. Finalmente, en la Sección 6 se presentan las conclusiones y posibles líneas de trabajo futuro.

2 Accesibilidad Web y Agentes Inteligentes

La Web puede ser un valioso recurso para los ciudadanos con capacidades diferentes y/o necesidades especiales. Las sociedades tienen el desafío de diseñar productos Web que posibiliten la percepción, comprensión e interacción a todos sus ciudadanos [9, 17]. La comunidad de expertos W3C/WAI ofrece recursos para enfrentar la accesibilidad de: (i) contenidos, (ii) navegadores, (iii) tecnologías de asistencia, (iv) software de desarrollo (*Authoring Tools*) y, (v) software de evaluación de accesibilidad (*Assessment Tools*). Por otra parte, el desarrollo de agentes inteligentes ha permitido asistir de diversas maneras al usuario Web. Un "agente" [8] es un sistema computacional que: (i) "tiene" objetivos, sensores y efectores, (ii) "decide" autónomamente y en tiempo real que acciones llevar adelante para maximizar el progreso hacia sus objetivos y, (iii) "aprende" y se "adapta" para mejorar su efectividad. A continuación, se propone una *Taxonomía de Agentes* para llevar adelante el proceso de revisión y facilitar la identificación y comprensión de los agentes inteligentes.

2.1 Una Taxonomía de Agentes Inteligentes

Una *Taxonomías de Agentes* proporciona conocimiento para facilitar la identificación y/o selección de los tipos de agentes de un sistema con rasgos inteligentes. A partir de la profunda revisión del área de interés a este trabajo, en la Fig. 1 se propone una *Taxonomía de Agentes* inteligentes, la cual se describe brevemente a continuación.

Las dos primeras ramas polarizan entre los agentes que operan individualmente y los que operan en conjunto con otros agentes. En la rama de los *Agentes Individuales*,

Maes [8] propone clasificar los agentes de software en *Agentes de Interfaz* y *Agentes de Tarea*. Mientras los primeros, asisten al usuario y pueden actuar en su nombre al conocer sus intereses, preferencias y hábitos; los segundos, ejecutan tareas para algunos usuarios y están personalizados a una tarea específica más que a una persona. Tomas & Fischer [16] proponen además un tercer tipo de Agentes Individuales denominados *Agentes de Red* (que conocen cómo comunicarse a través de la red) y además un sub-tipo dentro de los *Agentes de Interfaz* denominados *Agentes de Vista* (que activan la vista al usuario a partir de obtener la descripción de la vista del perfil de uso.) En la rama de los *Sistemas Multiagentes*, Wong & Sycara [19] presentan dos tipos de agentes: *Agentes Finales* ("*end-agents"*) y *Agentes Intermediarios* ("*middle-agents - MAs"*). Los primeros actúan de proveedores (cuando ofrecen servicios) y de solicitantes (cuando demandan servicios); mientras que los *MAs* permiten la interacción entre los *Agentes Finales*. Wong & Sycara [19] proponen una taxonomía que analiza el grado de intermediación entre los *Agentes Finales* e identifican los *Agentes Matchmaker* (que no auspician exclusivamente de intermediarios) y los *Agentes Facilitadores* (que actúa como únicos intermediadores entre los *Agentes Finales*.) En este punto, cabe señalar que Moya & Tolk [12] proponen la identificación de otros parámetros que aportan a los *Sistemas Multiagentes*, tales como el ambiente del agente (mecanismo de razonamiento y cooperación), la población del agente y las características del agente en la población. Finalmente, Huang et al.[5], propone una taxonomía de agentes Web para asistir a servidores (o a clientes) de realidad virtual de dos y tres dimensiones (2D y 3D). Tal como ilustra la Fig. 1, estos tipos de agentes se ubican en las hojas de nuestra taxonomía, ya sea asistiendo a los *Agentes Individuales* o a los *Sistemas Multiagentes* cooperando entre sí o auspiciando de intermediarios, ya sea entre otros agentes o con el usuario propiamente dicho.

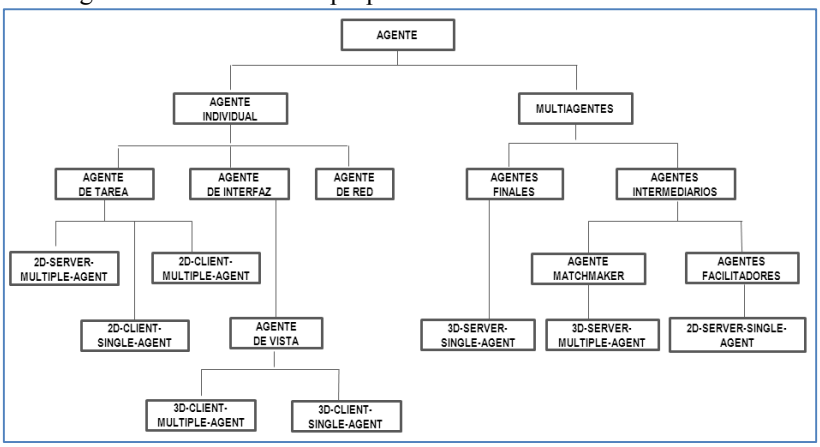


Fig. 1. Taxonomía de Agentes Inteligentes Propuesta.

3 Soporte Inteligente a la Accesibilidad Web

En base el estudio propuesto en Mazalu [10], se realiza una revisión actualizada y profunda para identificar aquellos enfoques que exhiben rasgos inteligentes y cuya

antigüedad no sea mayor a los 5(cinco) años. A continuación se presentan y describen brevemente los 8(ocho) enfoques seleccionados que incorporan el uso de agentes para dar soporte a la Accesibilidad Web.

Sistema de Filtrado de Información No Visual. Puzis [15] presenta un sistema Web denominado "Asistente Automatizado" que responde a un *Agente de Interfaz* para el filtrado de información no visual. El agente presenta sugerencias de contenido que pueden ser "consumidos" por el usuario, y acciones que pueden ser automatizadas en nombre del usuario. El *Agente de Interfaz* (asistente) puede: (i) pedirle al usuario que confirme una acción/es antes de la ejecución, (ii) pedirle al usuario que valide una acción/es luego de la ejecución o, (iii) actuar de forma totalmente autónoma. Este enfoque es potencialmente útil para usuarios con (o sin) discapacidad visual. El asistente utiliza un modelo predictivo para proporcionar sugerencias relevantes y minimizar la carga cognitiva del usuario. La característica más relevante del enfoque es la incorporación de la capacidad de filtrado a la navegación regular para: (i) reducir la gran cantidad de información que normalmente se presenta en las páginas Web y, (ii) enfocar a los usuarios en los contenidos más relevantes. El enfoque se integra en la implementación del "*screen-reader*" estándar (similar a JAWS[1]) y no especifica el uso de ninguna directriz de accesibilidad en particular.

Agentes para Asistencia de Discapacitados Visuales. Zhu et al.[22] presentan un agente basado en voz denominado *Sasayaki* que aumenta la salida de voz principal de un navegador por un segundo canal (voz secundaria), proporcionando información relevante al contexto/tarea o en respuesta a solicitudes del usuario. El agente *Sasayaki* asiste a usuarios con (o sin) discapacidad visual, realizando un seguimiento mediante la voz de la situación de la navegación y del comportamiento del usuario. Un prototipo de *Sasayaki* se implementó como un plug-in para el navegador de voz denominado *aiBrowser*[2], y pruebas piloto indican mejoras en los tiempos de búsqueda y en el nivel de confianza del usuario; se propone a futuro una mejora para incorporar las preferencias del usuario. El enfoque no especifica el uso de ninguna directriz de accesibilidad.

Agentes para Asistencia a Adultos Mayores. Chattaraman et al.[2] propone un agente virtual para asistir a usuarios de la tercera edad en la utilización de sitios de comercio electrónico. El enfoque provee soporte a la búsqueda y navegación pero además asiste al procedimiento de compra en línea. Este tipo de agente virtual no sólo se propone para asistir a los usuarios adultos mayores, sino que se sugiere su aplicación a otros dominios tales como *e-banking*, *e-helth* y *e-learning*. El enfoque no especifica el uso de ninguna directriz de accesibilidad en particular.

Agentes para la Identificación de Widgets. Chen et al.[3] proponen identificar los *widgets* a partir del código fuente mediante ingeniería inversa y evaluar la accesibilidad del contenido que manipula el *widget*. El agente aplica un método que identifica automáticamente el contenido dinámico (*widgets*) de una página Web. Para facilitar la definición e identificación de los *widgets* (los cuales muchas veces comparten elementos comunes), se propone una ontología como sistema de clasificación. La identificación de contenido Web dinámico, asiste a usuarios adultos mayores y con disca-

[1] Job Access With Speech <http://www.freedomscientific.com/jaws-hq.asp>
[2] Accessibility Internet Browser <http://www.eclipse.org/actf/downloads/tools/aiBrowser/>

pacidad visual, como así también puede asistir al desarrollador en la fase de diseño, ya que el agente puede sugerir el *mark-up* más adecuado para propiciar la accesibilidad. Lunn & Harper [7] proveen antecedentes a este enfoque proponiendo una herramienta denominada *SCWeb*[3] que utilizan expresiones lingüísticas y videos simples para explicar el funcionamiento del contenido dinámico. El enfoque no especifica el uso de ninguna directriz de accesibilidad en particular.

Herramienta de Refactorización para Mejorar la Accesibilidad y Usabilidad. Garrido et al.[4] proponen adaptar el concepto de refactorización para mejorar los atributos externos de una aplicación Web (usabilidad y accesibilidad.) Este enfoque provee una interfaz con posibles refactorizaciones (puntos de vista personalizados y accesibles de la aplicación) para mejorar la accesibilidad Web en el navegador del cliente. La propuesta, denominada *Client-Side Web Refactoring* (CSWR), permite crear automáticamente diferentes vistas personalizadas que mejoran el *look-and-feel*, aspectos de la estructura de navegación e interacción con la aplicación Web, preservando su funcionalidad. El proceso de refactorización remueve los *bad-smells* de usabilidad y accesibilidad, preservando el contenido y las operaciones. El CSWR se testeó con usuarios con discapacidad visual, aunque soluciones similares pueden aplicarse a otras discapacidades. El enfoque se implementa como un plug-in en la interfaz del navegador y propone compatibilidad con las directrices de la W3C.

Herramienta para Reportar Problemas de Accesibilidad. La accesibilidad en uso plantea problemas que no pueden ser resueltos solo con la conformidad a las directrices [14]. Vigo & Harper [18] proponen un método basado en la interacción con el usuario por medio de *WebTactics*, una herramienta que detecta y reporta problemas de accesibilidad a partir de la identificación de tácticas del usuario. La propuesta evalúa automáticamente la accesibilidad observando el uso del sitio Web para: (i) recopilar situaciones problemáticas que experimentan los usuarios con discapacidad visual e identificar las tácticas evasivas aplicadas, (ii) diseñar algoritmos que detecten automáticamente estos comportamientos (tácticas) para identificar problemas de accesibilidad y, (iii) aplicar los algoritmos para recolectar los problemas de accesibilidad. *WebTactics* se incorpora al navegador Mozilla Firefox e incluye el uso de agentes inteligentes. La propuesta se promueve para ser utilizada en conjunto con las directrices de la W3C.

Herramienta para Evaluación de Accesibilidad RIA. Doush et. al[1] presentan un framework conceptual para la evaluación automática de la accesibilidad de RIAs (*Rich Internet Applications*). El framework propuesto implementa un enfoque de evaluación donde a partir de un controlador y evaluador de eventos se analizan todos los componentes dinámicos. Luego, se genera un informe de síntesis de los elementos RIAs que presentan problemas de accesibilidad para ser inspeccionados por un experto. La herramienta se presenta como una extensión de un navegador Web. La propuesta considera las especificaciones WAI-ARIA para la evaluación de los componentes.

Herramienta para Evaluación de Accesibilidad. Mosqueira Rey et al.[11] proponen el framework GAEL que funciona sobre formatos HTML y CSS. La herramienta está disponible en línea para ser utilizada por desarrolladores y usuarios finales. La carac-

[3] SCWeb Assistant Tool <http://wel.cs.manchester.ac.uk/tools/extensions/scweb2/>

terística más relevante de la propuesta es la capacidad de aprendizaje operada por dos tipos de agentes inteligentes: (i) los agentes de usuario, que son capaces de navegar hasta un URL (de una página destino) a partir de un URL (de una página origen) y, (ii) los agentes analizadores de código HTML y CSS, que son capaces de inspeccionar el código de las páginas Web y extraer los datos que son útiles para el análisis. La propuesta aplica las recomendaciones WCAG[4] 2.0.

4 Nuestro Framework de Evaluación

El *Framework de Evaluación* está basado en el método de *Análisis de Características* (*Feature Analysis*), el cual esta definido dentro de la metodología DESMET[6]. El *Análisis de Características* es una evaluación cualitativa, basada en la identificación de requerimientos para una tarea o actividad particular y, el mapeo de dichos requerimientos a características que un enfoque debe proveer; en nuestro caso el enfoque debe dar soporte a la accesibilidad Web.

A continuación se proponen las 6(seis) características y las respectivas escalas de valoración, que utiliza nuestro *Framework de Evaluación* para efectuar el análisis de cada uno de los 8(ocho) enfoques seleccionados.

- **Tipo de Agente:** Se determina aplicando la *Taxonomía de Agentes* inteligentes propuesta en la Sección 2.1 (Tabla 1).

Tabla 1. Valoración para la característica Tipo de Agente.

TIPO DE AGENTE	VALORACIÓN (%)
AGENTE INDIVIDUAL (DE TAREA - DE INTERFAZ - DE RED)	50
MULTIAGENTES (FINALES - INTERMEDIARIOS)	75
COMBINACIÓN DE AMBOS	100

- **Ámbito.** Especifica la actividad de soporte que brinda el enfoque (Tabla 2).

Tabla 2. Valoración para la característica Ámbito.

ÁMBITO	VALORACIÓN (%)
EVALUACIÓN	25
EVALUACIÓN Y REPARACIÓN	50
FILTRADO Y TRANSFORMACIÓN	75
DESARROLLO	100

- **Directriz.** Indica la directriz de Accesibilidad que aplica el enfoque (Tabla 3).

Tabla 3. Escala de valores para la característica Directriz.

DIRECTRIZ	VALORACIÓN (%)
NINGUNA	0
WCAG 1.0 // ATAG 1.0	25
WCAG 2.0 // ATAG 2.0	50
WCAG 2.0 // 1.0 + ATAG 2.0 // 1.0	75
WCAG 2.0 // 1.0 + ATAG 2.0 // 1.0 + OTRAS	100

- **Ambiente de Ejecución.** Define la disponibilidad del enfoque y considera el tipo de usuario a la que está dirigida (Tabla 4).

[4] Web Content Accessibility Guidelines (1.0 / 2.0) <http://www.w3.org/WAI/intro/wcag.php>

Tabla 4. Valoración para la característica Ambiente de Ejecución.

AMBIENTE DE EJECUCIÓN	VALORACIÓN (%)
APLICACIÓN DE ESCRITORIO (DESARROLLADOR - USUARIO FINAL)	25
EN LÍNEA (DESARROLLADOR - USUARIO FINAL)	50
EXTENSIÓN DE NAVEGADOR (DESARROLLADOR - USUARIO FINAL)	75
EXTENSIÓN DE FRAMEWORK DE DESARROLLO (DESARROLLADOR)	100

- **Grado de Inteligencia.** Analiza los rasgos inteligentes presentes en el enfoque (Tabla 5).

Tabla 5. Valoración para la caracteristica Grado de Inteligencia.

GRADO DE INTELIGENCIA	VALORACIÓN (%)
USO DE HEURÍSTICAS (POR EJEMPLO: ANÁLISIS DEL CONTEXTO)	50
CAPACIDAD DE APRENDIZAJE	100

- **Barrera Física:** Determina la discapacidad en la cual el enfoque focaliza los esfuerzos para derribar barreras de accesibilidad (Tabla 6).

Tabla 6. Valoración para la característica Barrera Física.

BARRERA FÍSICA	VALORACIÓN (%)
AUDITIVA // TRASTORNOS EN EL HABLA Y LENGUAJE	25
PSICOMOTRIZ // COGNITIVA // NEUROLÓGICA	50
VISUAL // DETERIOROS ASOCIADOS AL ADULTO MAYOR	75
VISUAL + DETERIOROS ASOCIADOS AL ADULTO MAYOR + OTRAS	100

4.1 Aplicando el Framework

La Tabla 7 resume los resultados del proceso de evaluación de los 8(ocho) enfoques de acuerdo al *Análisis de Características* que propone el *Framework de Evaluación*.

Tabla 7. Resumen de los Resultados del Proceso de Evaluación

ENFOQUE	TIPO DE AGENTE	ÁMBITO	DIRECTRIZ	AMBIENTE DE EJECUCIÓN	GRADO DE INTELIGENCIA	BARRERA FÍSICA
PUZIS[15]	50	75	0	25	100	75
ZHU ET AL.[22]	50	75	0	75	50	75
CHATTARAMAN ET AL.[2]	50	75	0	75	50	75
LUNN & HARPER [7]	50	75	0	75	50	75
GARRIDO ET AL.[4]	50	75	100	75	50	75
VIGO & HARPER[18]	75	25	0	75	50	75
DOUSH ET AL[1]	75	25	100	75	50	75
MOSQUEIRA REY ET AL.[11]	75	25	75	50	100	75

Una primera lectura de los resultados (Tabla 7) nos permite determinar que: (i) el **Tipo de Agente** (Fig. 2) más utilizado es el agente individual de interfaz y de tarea, ya que lo utilizan 5(cinco) enfoques; (ii) el **Ámbito** (Fig. 3) preferido para propiciar la accesibilidad es aplicar filtrado y transformación de contenido, ya que lo seleccionan 5(cinco) enfoques y; (iii) el **Grado de Inteligencia** (Fig. 4) en general se remite a usar algún tipo de heurística, tal como el análisis del contexto, ya que sólo 2(dos) enfoques exhiben rasgos con capacidad de aprendizaje.

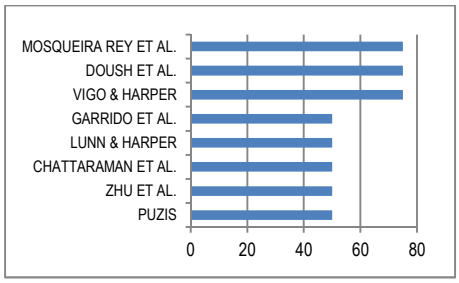


Fig. 2. Valoración de la característica Tipo de Agente.

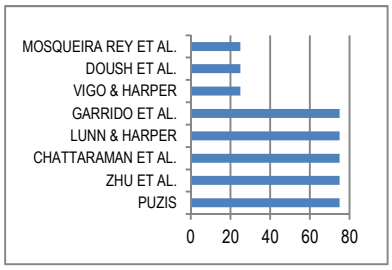


Fig. 3. Valoración de la característica Ámbito.

Con respecto a las características **Directriz**, **Ambiente de Ejecución** y **Barrera Física**: (i) sólo 3(tres) enfoques consideran especificaciones referentes de accesibilidad; (ii) 6(seis) enfoques prefieren operar como una extensión o *plug-in* del navegador Web y; (iii) los 8(ocho) enfoques están focalizados en la discapacidad visual y/o deterioros físicos asociados al adulto mayor.

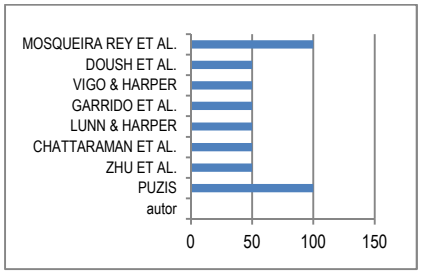


Fig. 4. Valoración de la característica Grado de Inteligencia.

5 Discusión

La *Taxonomía de Agentes* (Fig. 1), resume el desarrollo de agentes inteligentes Web de los últimos años [12, 16, 19]. Particularmente la Tabla 7, ilustra a través de sus resultados el desarrollo de los agentes de accesibilidad Web. La revisión y evaluación del estado del arte revela que un 62 % de los enfoques aplican *Agentes de Interfaz* y *Agentes de Tarea* para minimizar las barreras de accesibilidad a usuarios con discapacidad visual y a usuarios que acceden desde algún dispositivo móvil [20]. Estas preferencias en el desarrollo de agentes se debe a que: (i) estos tipos de agentes asisten

directamente a la interacción y navegación del usuario y; (ii) derribar barreras asociadas a la discapacidad visual favorece también el acceso a otros usuarios con (y sin) discapacidad. Respecto a la interacción usuario-agente, las tareas que propicia son: (i) dar soporte a operaciones en línea, por ejemplo del tipo *e-commerce* que propone el enfoque de Chattaraman et al.[2]; (ii) asistir en el uso de herramientas de comunicación, por ejemplo del tipo correo electrónico que propone el enfoque de Garrido et al.[4] y; (iii) facilitar la navegación de aplicaciones con estructuras complejas y contenido dinámico, por ejemplo del tipo RIAs que propone el enfoque de Chen et al.[3].

Propiciar la accesibilidad Web, dando buen soporte tanto del lado del desarrollador como del usuario final, es un procedimiento complejo que requiere de la participación y trabajo colaborativo de un conjunto de agentes. Los sistemas multiagentes, que combinen las fortalezas de los agentes individuales, no se han desarrollado en su máximo potencial ni aplicado en forma exhaustiva para dar soporte al usuario en la interacción y navegación. Esta es un área de trabajo que aún requiere ser explorada y desarrollada en futuras investigaciones. La construcción de perfiles de usuario, su vinculación con tipos de aplicaciones Web y el uso de repositorios accesibles, resulta una combinación de componentes que los agentes pueden orquestar con destreza para proveer soporte inteligente y facilitar verdaderamente la accesibilidad a la Web.

Las características particulares de los agentes inteligentes tales como autonomía, movilidad, adaptabilidad e iniciativa, definitivamente pueden contribuir en: (i) la evaluación de productos Web, donde la intervención y juicio humano para la detección de falsos positivos es aún causa de sobrecarga y demora en la corrección de fallas, (ii) el desarrollo de herramientas de autor que apliquen las directrices ATAG[5] y; (iii) la transformación y filtrado de contenido en sitios Web existentes.

6 Conclusiones y Trabajos Futuros

Para asistir a la identificación y comprensión de los diferentes tipos de agentes, se propone una *Taxonomía de Agentes* inteligentes. Con esta taxonomía como referente, el objetivo de este trabajo se enfoca en: (i) realizar una revisión profunda del estado del arte de los enfoques de accesibilidad que exhiben rasgos inteligentes y; (ii) desarrollar y aplicar un *Framework de Evaluación* para analizar y determinar fortalezas y debilidades. Los resultados confirman que el grado de evolución alcanzado por el estado del arte no es suficiente para satisfacer las necesidades de accesibilidad del usuario Web. Profundizar sobre las posibilidades que ofrecen los sistemas multiagentes, nos permitirá reforzar y encausar el trabajo futuro para proveer soporte inteligente y automático que propicie la accesibilidad del lado del usuario y del desarrollador.

Agradecimientos. Este trabajo es soportado por el Proyecto UNPA 29/B144 “Diseño y Evaluación de Portales Web” y en colaboración con el Proyecto UNComa 04/F001 “Reuso Orientado a Dominios”, bajo el programa “Desarrollo Basado en Reuso”.

Referencias

1. Abu Doush, I., et al. *The design of RIA accessibility evaluation tool*. in *Advances in Engineering Software*. 2013: Elsevier.

[5] Authoring Tool Accessibility Guidelines (1.0 / 2.0) <http://www.w3.org/WAI/intro/atag.php>

2. Chattaraman, V., W.S. Kwon, and J.E. Gilbert, *Virtual agents in retail web sites: Benefits of simulated social interaction for older users.* Computers in Human Behavior, 2012.
3. Chen, A.Q., et al., *Widget Identification: A High-Level Approach to Accessibility.* World Wide Web, 2012: p. 1-17.
4. Garrido, A., et al., *Personalized Web Accessibility Using Client-Side Refactoring.* 2012.
5. Huang, Z., et al. *A taxonomy of web agents*. in *Database and Expert Systems Applications, 2000. Proceedings. 11th International Workshop on*. 2000: IEEE.
6. Kitchenham, B., S. Linkman, and D. Law, *DESMET: a methodology for evaluating software engineering methods and tools.* Computing & Control Engineering Journal, 1997. **8**(3): p. 120-126.
7. Lunn, D. and S. Harper, *Providing assistance to older users of dynamic Web content.* Computers in Human Behavior, 2011. **27**(6): p. 2098-2107.
8. Maes, P., *InteIligent Software.* Scienttjic American, 1995. **Vol. 273, No.3**: p. 84-86.
9. Martín, A., A. Cechich, and G. Rossi. *Accessibility at early stages: insights from the designer perspective*. in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. 2011: ACM.
10. Mazalu, A.C.A.Z.R., *Utilización de Técnicas Inteligentes en el Soporte a la Accesibilidad Web*, in *Facultad de Ciencias Exactas*. 2012, Universidad Nacional del Centro de la Provincia de Buenos Aires(UNICEN): Tandil.
11. Mosqueira-Rey, E., et al., *A multi-agent system based on evolutionary learning for the usability analysis of websites*, in *Intelligent Agents in the Evolution of Web and Applications*. 2009, Springer. p. 11-34.
12. Moya, L.J. and A. Tolk. *Towards a taxonomy of agents and multi-agent systems*. in *Proceedings of the 2007 spring simulation multiconference-Volume 2*. 2007: Society for Computer Simulation International.
13. Peredo, R., et al., *Intelligent Web-based education system for adaptive learning.* Expert Systems with Applications, 2011. **38**(12): p. 14690-14702.
14. Power, C., et al. *Guidelines are only half of the story: accessibility problems encountered by blind users on the web*. in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. 2012: ACM.
15. Puzis, Y. *An interface agent for non-visual, accessible web automation*. in *Adjunct proceedings of the 25th annual ACM symposium on User interface software and technology*. 2012: ACM.
16. Thomas, C.G. and G. Fischer. *Using agents to improve the usability and usefulness of the World-Wide Web*. in *Fifth International Conference on User Modeling*. 1996.
17. Trewin, S., et al. *Accessibility challenges and tool features: an IBM Web developer perspective*. in *Proceedings of the 2010 international cross disciplinary conference on web accessibility (W4A)*. 2010: ACM.
18. Vigo, M. and S. Harper. *Evaluating accessibility-in-use*. in *Proceedings of the 2013 International Cross Disciplinary Conference on Web Accessibility (W4A), W4A*. 2013.
19. Wong, H.C. and K. Sycara. *A taxonomy of middle-agents for the internet*. in *Proc. 4th Int'l Conf. Multiagent Systems*. 2000.
20. Yesilada, Y., G. Brajnik, and S. Harper, *Barriers common to mobile and disabled web users.* Interacting with Computers, 2011. **23**(5): p. 525-542.
21. Zhou, L., A.S. Mohammed, and D. Zhang, *Mobile personal information management agent: Supporting natural language interface and application integration.* Information Processing & Management, 2012. **48**(1): p. 23-31.
22. Zhu, S., et al. *Sasayaki: an augmented voice-based web browsing experience*. in *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 2010: ACM.

Evaluación de Accesibilidad del Contenido Web Utilizando Agentes

Rafaela Mazalu[12], Alejandra Cechich[1], and Adriana Martín[13]

[1] Giisco, Facultad de Informática, Universidad Nacional del Comahue, Neuquén, Argentina
[2] Concejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Comahue, Neuquén, Argentina
[3] Unidad Académica Caleta Olivia, Universidad Nacional de la Patagonia Austral, Caleta Olivia, Santa Cruz, Argentina,
rafaelamazalu@gmail.com,acechich@gmail.com,adrianaelba.martin@gmail.com

Abstract. Actualmente, existe un creciente número de herramientas que permiten a los desarrolladores Web evaluar la accesibilidad de sus páginas y sitios Web. Muchas herramientas también sugieren al desarrollador realizar reparaciones específicas; y algunas herramientas siguen automáticamente los enlaces para evaluar múltiples páginas dentro de un sitio o un dominio completo. Aunque este tipo de herramientas pueden resultar muy útiles en la identificación de problemas de accesibilidad, muchos problemas de accesibilidad son subjetivos y no pueden evaluarse sin una inspección manual. Nuestro enfoque está dirigido a la evaluación y reparación de la accesibilidad como actividades relacionadas que deben ser soportadas de manera inteligente y en forma automática. Para ello, se deben considerar varios aspectos, desde la identificación automática de las discapacidades de los usuarios a la reparación en si misma. En este artículo, introducimos una solucion basada en agentes para hacer frente al paso de evaluación de este enfoque, que esta basado en la identificación de aquellas barreras de accesibilidad presentes en el sitio que el usuario está navegando y que están relacionadas con las discapacidades visuales que el usuario posee. Además el procedimiento de evaluación se ilustra a través de un caso motivacional.

Keywords: Discapacidades Visuales, Agente Inteligente, Barreras de Accesibilidad, Accesibilidad Web

1 Introducción

La importancia de identificar las barreras de Accesibilidad de forma automática acorde al perfil del usuario puede ser significativa dado que la amplia variedad de aplicaciones Web no se encuentran organizadas según las necesidades de los usuarios. De está manera se excluye a una gran cantidad de usuarios con limitaciones como discapacidades, limitaciones con respecto al contexto de acceso, software, hardware, ancho de banda de la conexión, etc. [1].
Particularmente, la accesibilidad Web se refiere a la practica inclusiva de crear

páginas web utilizables por todas las personas. Para ello, muchos trabajos y enfoques [2, 8, 10, 12, 13, 18, 19] toman como referencia las Guías de Accesibilidad al Contenido Web 1.0 [3] y 2.0 [4].
Actualmente, existen algunos esfuerzos hacia la automatización de los aspectos de la accesibilidad Web. Por ejemplo, la inspección se puede automatizar mediante el uso de sistemas capaces de analizar y recomendar, como TAW [5], Bobby [2], y WAVE [6]. Estos sistemas tienen como objetivo evaluar las fortalezas y debilidades de los sitios Web, y se focalizan en ayudar a los diseñadores a mejorar la accesibilidad Web. También existen enfoques, que permiten evaluar automáticamente los formatos de los documentos Web, por ejemplo mediante el uso de GAEL [7] aprovechando sus capacidades de razonamiento.
Al enfocarnos en la evaluación y reparación automática, consideramos que las herramientas de soporte deberían ser inteligentes, en cuanto a adaptarse a las limitaciones individuales y a la situación actual de cada persona, para ofrecer un servicio lo más apropiado posible en relación con las intenciones y objetivos del usuario. La presencia de técnicas basadas en conceptos de inteligencia artificial son deseables en herramientas de reparación y transformación debido a la necesidad de simplificar el proceso de toma de decisiones, reduciendo así la intervención humana. Sin embargo, la limitada existencia de inteligencia en las herramientas actuales [7–11], hace que el usuario deba decidir sobre el proceso de operación.
La mayoría de las herramientas de evaluación y reparación existentes adhieren a los estándares de accesibilidad Web y soportan diferentes formatos de documentos Web. Sin embargo, existe una brecha entre el soporte existente y las necesidades reales en términos de detección automática e inteligente de barreras de accesibilidad Web. Este hecho, nos conduce a enfocarnos en resolver este tipo de barreras, para identificar y responder las necesidades particulares de los usuarios aplicando características inteligentes. Para ello, elegimos una solución basada en agentes. En un primer paso, nuestra meta fue diseñar un agente que crea y clasifica perfiles de usuarios en términos de estereotipos definidos según las limitaciones visuales y de contexto de uso que los usuarios puedan tener [20]. En la presente etapa, presentamos un agente cuya meta es proporcionar información sobre las barreras de accesibilidad presentes en la página web que el usuario está navegando de acuerdo a las características del perfil del usuario.
Este artículo se organiza de la siguiente manera. En la Sección 2 se introduce una solución basada en agentes que evalúa y corrige potenciales barreras de accesibilidad presentes en una página de acuerdo con las caracteísticas del perfil del usuario que navega la misma. Luego, en la Sección 3 definimos el agente evaluador de la propuesta mediante la descripción de sus objetivos, sus conocimientos, que incluyen las potenciales barreras de accesibilidad, y como estos son utilizados para la realización de la evaluación de paginas Web. En la Sección 4 ilustramos nuestra propuesta con un caso motivacional. Finalmente, en la Sección 5 presentamos las conclusiones y trabajo futuro.

2 Evaluación Automática de Barreras Utilizando Agentes

El proceso de evaluación de la accesibilidad Web tiene por objetivo analizar, estudiar y validar las páginas Web para que las mismas no presenten problemas de accesibilidad y cumplan con las pautas y directrices de accesibilidad web existentes. La evaluación de accesibilidad se puede realizar tanto en forma manual como automática. La primer forma requiere la participación de un experto para el análisis; mientras que en la segunda el análisis es realizado por una herramienta, que automatiza la tarea. Sin embargo, una revisión de este tipo cuenta con algunas limitaciones importantes, como por ejemplo puede no detectar errores o señalar errores que realmente no existen (falsos positivos).

La solución propuesta por nuestro enfoque es la mejora de la automatización mediante la utilización de un sistema multiagente que realiza el análisis, la evaluación y corrección de una página Web de manera inteligente. La solución

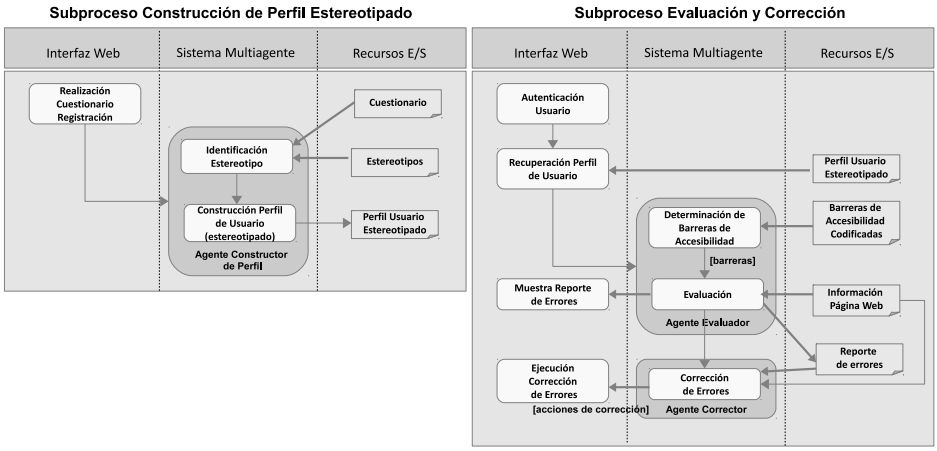


Fig. 1. Proceso marco para el sistema multiagente

propuesta es una herramienta Web que cumple su funcionalidad teniendo en cuenta un proceso marco que define la forma de interacción de sus componentes y la intervención del usuario. Dicho proceso esta conformado por dos actividades principales, las cuales son la *Identificación y Reconocimiento del Usuario* y *Análisis, Evaluación y Corrección de Páginas Web*. La primera de las actividades se encuentra relacionada con la identificación y clasificación del usuario de la herramienta, y es llevada a cabo mediante el subproceso *Construcción del Perfil Estereotipado* [20]. De la misma manera, la segunda actividad es llevada a cabo mediante el subproceso *Evaluación y Corrección*. La Fig. 1 muestra los subprocesos realizados en el marco de nuestra propuesta. El primer subproceso es realizado una sola vez para cada usuario. En el mismo, a través del cuestionario de registración completado por el usuario, el agente constructor de perfiles iden-

tifica el estereotipo al que pertenece, es decir, a que grupo de usuarios con determinadas características de discapacidad visual pertenece (Usuarios ciegos, disminuídos visuales, daltónicos, con epilepsia fotosensibles o sin discapacidad visual). Luego el agente construye el perfil del usuario que será una pieza clave en el resto del proceso [20]. El segundo subproceso involucra la evaluación y corrección de una página web en base al perfil del usuario generado anteriormente, y las posibles barreras de accesibilidad que pueden presentarse para el estereotipo del perfil. Este subproceso tendrá lugar al momento del análisis de una página Web por parte de un usuario ya registrado. Cabe destacar que la interacción entre el sistema multiagente y el usuario es llevada a cabo mediante una aplicación Web, que cuenta con la funcionalidad necesaria para la realización del cuestionario y para servir como forma de acceso y comunicación entre el sistema multiagente y el componente cliente de la herramienta.
Para la etapa de evaluación, principal objetivo de este trabajo, el sistema multiagente cuenta con un agente deliberativo, encargado de considerar cada barrera de accesibilidad que puede estar presente en la página Web evaluada de acuerdo a la discapacidad del usuario. El resultado de dicha evalución es un reporte con los errores de accesibilidad Web que presenta la página. Este reporte es procesado por el componente cliente de la herramienta y mostrado al usuario. Luego, para la etapa de corrección, otro agente deliberativo toma los errores de accesibilidad detectados y la información sobre la página analizada para generar el script de corrección de errores. Finalmente, el script de correcciones es interpretado y aplicado por el componente cliente en la página Web analizada.

3 Definiendo el Agente Evaluador

Un agente inteligente define un sistema basado en el conocimiento que percibe de su entorno [15], razona para interpretar sus percepciones, inferir y resolver problemas. Luego el agente define las acciones y tareas a realizar sobre el entorno para alcanzar el conjunto de objetivos para los que fue diseñado. La estructura del agente evaluador es deliberativa, basada en el modelo BDI (creencia, deseo e intención).

El entorno que percibe el mismo esta compuesto por los restantes agentes del sistema, la página Web que el usuario está navegando/evaluando y el perfil del usuario. Las creencias del agente incluyen el conocimiento que el agente tiene de si mismo y de su entorno. En este caso, las creencias incluyen el perfil del usuario, la página Web bajo evaluación y las posibles barreras de accesibilidad que se pueden presentar. Los deseos son las metas que el agente quiere cumplir a largo plazo. En nuestro caso, incluimos como objetivo identificar las barreras de accesibilidad web que presenta la página bajo evaluación de acuerdo al perfil del usuario, es decir, acorde a las limitaciones visuales que el usuario presenta, e identificando que puntos de verificación de las WCAG 1.0 y criterios de éxito de las WCAG 2.0 corresponde al caso. Las intenciones son los objetivos que el agente intenta lograr. Para este agente, el objetivo es descubrir la mayor cantidad posible de barreras de accesibilidad presentes en la página bajo evaluación.

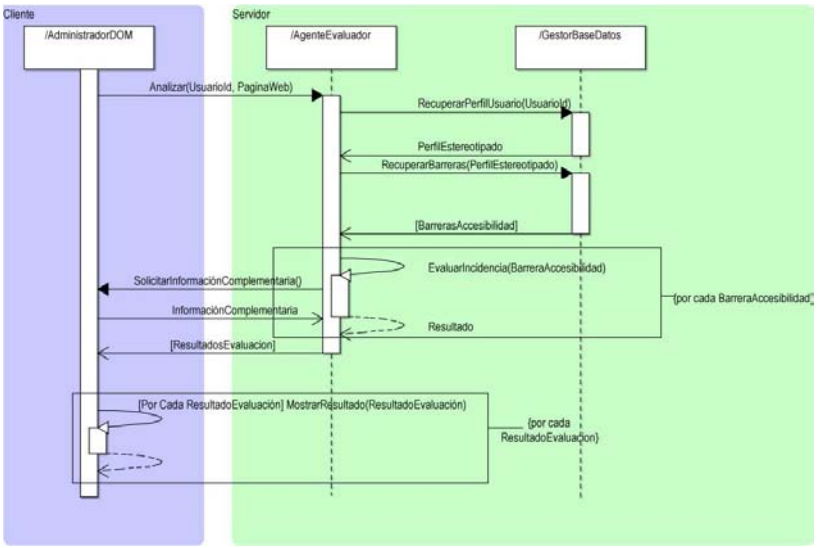


Fig. 2. Diagrama de secuencia de la interacción del agente evaluador con su entorno de trabajo

Dado que este agente se basa en razonamiento práctico [16], su decisión en cada acción que realiza le facilita arribar a sus objetivos. Para ilustrar la relación del agente evaluador con su entorno y las interacciones que realiza para alcanzar sus objetivos, utilizamos el diagrama de secuencia que se visualiza en la Fig. 2. En este diagrama podemos observar que las peticiones de análisis enviadas por el *AdministradorDOM* para un usuario y página Web particulares. En respuesta a esta solicitud, el agente interactúa con el *GestorBaseDatos* para recuperar el perfil estereotipado del usuario [20].

Una vez recuperado el perfil, el agente determina bajo qué estereotipo fue clasificado y nuevamente interactúa con el GestorBaseDatos para obtener todas las posibles barreras que se pueden presentar para este tipo de estereotipo. La representación de dichas barreras está basada en el mapeo definido en la propuesta de Bustos et al. [21]. Luego, por cada una de las potenciales barreras, el agente evalúa las incidencias de las mismas en la página bajo análisis. El agente evaluador puede solicitar al AdministradorDOM información adicional para enriquecer el proceso de análisis de cada barrera y arribar a una conclusión con precisión. Una vez concluido el proceso de evaluación devuelve el resultado del mismo de una forma que el AdministradorDOM puede interpretar y mostrar al usuario final.

3.1 Identificando Barreras de Accesibilidad

En nuestro proceso de identificación de posibles barreras de accesibilidad, primeramente, analizaremos aquellas que normalmente pueden encontrarse en un sitio

Web, y que dificultan la interacción de los usuarios con discapacidades visuales. Luego indicaremos los puntos de verificación presentes en las WCAG 1.0 [3] y los criterios de éxito de las WCAG 2.0 [4] que tienen asociados cada una de estas barreras y las posibles soluciones a dichas barreras. Para alcanzar este objetivo, consideramos el enfoque "BarrierWalkthrough" de Giorgio Brajnik [14]. El mismo consiste en que un evaluador tiene que considerar un número predefinido de posibles barreras, las cuales son interpretaciones y extensiones de los principios de Accesibilidad [3, 4, 17]. El contexto para la aplicación de este método comprende las distintas categorías de usuarios (como usuario ciegos, disminuídos visuales, daltónicos y fotosensibles), los escenarios de uso de los sitios Web (como uso de un lector de pantalla), y los objetivos propios de usuario.

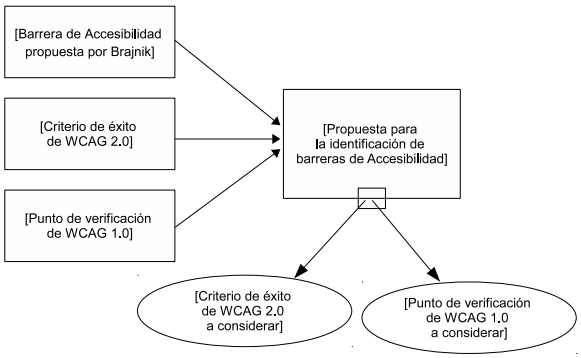


Fig. 3. Propuesta identificación de puntos de Verificación correspondientes a las WCAG para evaluar sitios Web para usuarios con discapacidad visual

Para analizar, seleccionar y reorganizar las barreras de accesibilidad junto con los puntos de verificación de las WCAG, consideramos lo siguiente: (i) El obstáculo o barrera de accesibilidad que impide realizar la tarea o actividad al usuario; (ii) La forma en que los elementos u objetos son innaccesibles; (iii) Las posibles soluciones que deben aplicarse para que desaparezca el obstáculo.

Para plasmar nuestra propuesta de identificación de barreras utilizaremos el siguiente proceso de mapeo y la plantilla de resultados. Como puede observarse en la Fig. 3 para el proceso de mapeo consideramos cada una de las barreras de accesibilidad propuestas por Brajnik [14] y cotejamos con ellas los puntos de verificación de las WCAG 1.0 y criterios de éxito de las WCAG 2.0. En caso de que el objetivo del punto de verificación (y criterios de éxito) se encuentre incluído dentro de la descripción de la barrera de accesibilidad, entonces dicho punto de verificación (y criterio de éxito) será considerado al momento de realizar una evaluación para la detección de la barrera en un sitio Web.

En la plantilla para la identificación de barreras de accesibilidad que se muestra en la Tabla 1 se especifican: (i) la causa que provoca la presencia de la barrera; (ii) las consecuencias de dicha presencia; (iii) posibles soluciones que pueden apli-

Table 1. Plantilla para la identificación de Barreras de Accesibilidad

Causa	Descripción de la barrera de Accesibilidad que se encuentra en el sitio Web
Consecuencia	Son los posibles problemas que afectarán los objetivos de los usuarios cuando se encuentren con la barrera al navegar por el sitio Web
Solución	Las acciones o cambios que se deben realizar para que la barrera desaparezca
Criterios de Éxito WCAG 2.0:	Criterios de éxito de las WCAG 2.0 que estan incluídos en la problemática que presenta la barrera.
Puntos de Verificación WCAG 1.0:	Puntos de verificación de las WCAG 1.0 que estan incluídos en la problemática que presenta la barrera.

carse para eliminar dicha barrera; y los puntos de verificación de las WCAG 1.0 y los criterios de éxito de las WCAG 2.0 que se encuentran incluídos en la barrera.

3.2 Aplicando la Propuesta de Identificación de Barreras

Para describir la propuesta de identificación de barreras de Accesibilidad tomamos como ejemplo la barrera "*Imágenes Ricas que carecen de un texto equivalente*" que afecta a usuarios con ceguera.
Causa. La página contiene alguna imagen que proporciona información (por ejemplo, un diagrama, histograma, imagen, dibujo, gráfico), pero sólo en un formato gráfico, no hay una descripción equivalente textual que aparezca en la página.
Consecuencia. El usuario no puede utilizar la información transmitida por la imagen. Reducción significativa de la eficacia y la productividad del usuario. El usuario, incluso si percibe que hay una imagen importante, no tiene manera de obtener la información que la misma contiene. Además, el usuario invierte tiempo y esfuerzo extra para averiguar en que parte de la página (o sitio) puede obtener información sobre lo que la imagen pretende trasmitir.
Solución. Añadir una descripción textual equivalente a la imagen mediante el atributo ALT o el atributo LONGDESC de IMG, y si no es suficiente con la etiqueta OBJECT, especificando el texto en el contenido de la etiqueta. Si esto todavía resulta insuficiente, se puede añadir un enlace en la imagen que lleve a una página específica donde la descripción textual esté presente. Otra estrategia es colocar el texto equivalente cerca de la imagen de manera que también pueda ser visto por aquellos que pueden ver en la imagen.
Criterios de Éxito WCAG 2.0. 1.1: 1.1.1
Puntos de Verificación WCAG 1.0. 1.1

4 Caso Motivacional

Para mostrar el funcionamiento del agente evaluador de una forma práctica utilizamos el sitio Web de la Facultad de Informática de la Universidad Nacional del

Comahue[4]. El mismo está destinado a jóvenes y adultos interesados en obtener información sobre dicha facultad, tal como información institucional, académica, de investigación y de extensión. En la Fig. 4 se visualiza la página de inicio del mismo, la cual en primera instancia ya cuenta con algunos inconvenientes que afectan a la accesibilidad, tal es el caso de las imágenes sin texto alternativo como las que se señalan en la misma imagen y cuyo código HTML se muestra resaltado.

Fig. 4. Captura de página de inicio del Sitio Web de la Facultad de Informática de la Universidad Nacional del Comahue

Para este caso, suponemos que un usuario con ceguera visual accede al sitio, al encontrarse con algunos inconvenientes en la navegación de la página solicita la asistencia de nuestra herramienta. Habiéndose registrado previamente, la herramienta recupera el perfil del usuario mediante un nombre y contraseña. La siguiente tarea a realizarse es el análisis de la página Web, para lo cual el agente evaluador de la herramienta recupera las potenciales barreras de accesibilidad para el estereotipo de usuarios ciegos al que pertenece el perfil del usuario. Dentro de estas barreras se encuentra las de imágenes ricas que carecen de texto equivalente. Ya en el proceso de evaluación de las barreras el agente detecta la presencia de imágenes con las características mencionadas, tales como las imágenes con el logo de la Universidad[5], el logo de la Plataforma para Educación a Distancia[6], el logo del sistema de gestión de alumnos que utiliza la Facultad[7], utilizadas como enlaces para acceso a los respectivos sitios. Ante esta situación el agente genera un reporte de error, tanto para el componente cliente, que muestra al usuario esta condición, como para el agente de corrección que intentará encontrar una descripción alternativa para cada imagen con el fin de subsanar la presencia de esta barrera. Cabe destacar que el agente evaluador no solo detecta la ausencia de un texto alternativo, de la misma manera trata de detectar textos alterna-

[4] http://faiweb.uncoma.edu.ar/
[5] UNComa: http://www.uncoma.edu.ar
[6] PEDCO: http://pedco.uncoma.edu.ar
[7] Siu Guaraní: http://siufai.uncoma.edu.ar/

tivos que carecen de significado para el usuario, tal como simbolos, nombre del archivo que contiene a la imagen, etc. El agente evaluador ha sido implementado en Spade[8], un framework para sistemas multiagentes basado en python, compatible con el estandart FIPA[9], independiente del lenguaje y plataforma. Las barreras son expresadas en lenguaje XML, lenguaje con el cual podemos modelar y estructurar con mayor flexibilidad la información pertinente a las barreras de accesibilidad, sin perder capacidad de procesamiento por parte de la herramienta.

5 Conclusiones y trabajo futuro

El desarrollo de aplicaciones Web accesibles es un factor fundamental para la concreción del principio básico de acceso universal. Sin embargo, la gran mayoría de las páginas en la Web ha sido desarrolladas desconociendo a los potenciales beneficiarios de un desarrollo accesible. Motivados por esta realidad, presentamos en este trabajo una solución basada en agentes inteligentes para la evaluación automática de barreras de accesibilidad. Para ello, partimos de las necesidades del usuario, sus discapacidades (si las presenta) y las barreras que puede llegar a encontrarse al navegar en la Web. Primeramente identificamos y vinculamos posibles barreras de accesibilidad a las WCAG 1.0 y 2.0. Luego, describimos el proceso marco que abarca toda nuestra propuesta. Seguidamente, definimos las metas, las intenciones y objetivos del agente evaluador en el que se enfoca el presente trabajo, así como el ambiente donde el mismo interactúa para alcanzar sus objetivos y, describimos su funcionalidad a través de un caso motivacional. Sin embargo, resta validar experimentalmente la propuesta así como extenderla para relacionar la evaluación a un proceso de reparación automática. Nuestros esfuerzos actuales se dirigen a estos aspectos, con el objetivo de facilitar el proceso de hacer la Web accesible para todos.

Agradecimientos Este trabajo es parcialmente soportado por el proyecto UNComa 04/F001 “Reuso Orientado a Dominios”, bajo el programa “Desarrollo Basado en Reuso”, y el proyecto PAE-PICT 2312.

References

1. Hassan Montero, Y., Martín Fernández, F.: Qué es la accesibilidad web. Electronic Magazine No Solo Usabilidad, 2003.
2. Martín, A., Cechich, A. and Rossi, G.: Comparing Approaches to Web Accessibility Assessment. Calero, C., Moraga, M A., Piattini, M. (eds.) Handbook of Research on Web Information Systems Quality, pp. 181-205. Information Science Reference, Hershey New York, 2008.
3. Word Wide Web Consortium (W3C): Web Content Accessibility Guidelines (WCAG) 1.0. Technical report, 1999.

[8] https://github.com/javipalanca/spade
[9] http://www.fipa.org/

4. Word Wide Web Consortium (W3C). Web Content Accessibility Guidelines (WCAG) 2.0. Technical report, 2008.
5. Centro para el Desarrollo de Tecnologías de la Información y Comunicación: Taw servicios de accesibilidad y movilidad web, 2013.
6. Web Accessibility in Mind (WebAIM): Wave wave is a free web accessibility evaluation tool, 2001.
7. Mosqueira Rey, E., Ríos, D. and Vázquez García, A.: Intelligent Agents in the Evolution of Web and Applications, volume 117 of ISBN: 978-3-540-88070-7, chapter A Multi-agent System Based on Evolutionary Learning for the Usability Analysis of Websites, pages 1137. Springer, 2009.
8. Bigham, J., Kaminsky, R., Ladner, R. , Danielsson, O. and Hempton, G. :Webinsight: Making web images accessible. In The 8th international ACM Conference on Assistive Technologies - ASSETS, pages 181188, 2006.
9. SSB BART Group: Infocus quick reference amp. Recovery in 2013, at https://www.ssbbartgroup.com/reference/index.php/InFocus_Quick_Reference.
10. Di Lucca, G., Fasolino, A. and Tramontana, P.: Web site accessibility: Identifying and fixing accessibility problems. In Seventh IEEE International Symposium on Web Site Evolution WSE05, pages 7178. ACM, 2005.
11. Pontelli, E., Son, T., Kottapall, K., Ngo, C., Reddy, R. and Gillan, D.: A system for automatic structure discovery and reasoning-based navigation of the web. Interacting with Computers, volume 16, pages 451475. Elsevier, 2004.
12. Keates, S. and Clarkson, P.: Countering design exclusion: bridging the gap between usability and accessibility. Universal Access in Information Society, 2:pp215255, 2003.
13. Yesilada, Y., Harper,S., Goble, G. and Stevens, R.: Screen readers cannot see: Ontology based semantic annotation for visually impaired web travelers. In Proceedings of the International Conference on Web Engineering (ICWE2004), pages 445458, 2004.
14. Brajnik, G.: Barrier walkthrough - heuristic evaluation guided by accessibility barriers. Recovery in 2013, at http://sole.dimi.uniud.it/ giorgio.brajnik/projects/bw/bw.html, 2009.
15. Russell, S. and Norvig, P.: Artificial Intelligence: A Modern Approach, Prentice-Hall, Englewood Cliffs, NJ, 1995.
16. Wooldridge, M.: An Introduction to Multiagent Systems. Department of Computer Science, University of Liverpool, UK, 2002.
17. Government U. F.: Quick reference guide to section 508 resource documents. Published in http://www.section508.va.gov/, 2003.
18. Gibson, B.: Enabling an accessible web 2.0, volume 225, pages 16. ACM, New York, 2007.
19. Matera, M., Rizzo, F. and Carughi, G.: Web usability: Principles and evaluation methods. In Emilia Mendes and Nile Mosley, editors, Web Engineering, pages 143180. Springer Berlin Heidelberg, 2006.
20. Mazalu, R., Cechich, A. and Martín, A.: Automatic Profile Generation for Visual-Impaired Users. Aceptado para publicación en el 14 Simposio Argentino en Ingeniería de Software, Argentina, 2013.
21. Bustos, B., Martín, A. and Cechich, A.:Diseño de Interfaces Guiado por Restricciones de Accesibilidad Web. XIII Congreso Iberoamericano en "Software Enginnering", Ecuador, 2010.

QUCO2: Development of a tool for measuring the quality of Web applications

Nicolás Tortosa[1], Noelia Pinto[1], César Acuña[1], Liliana Cuenca Pletsch[1], Marcelo Estayno[2]

[1]Software Quality Research Group, (GICS). Department of Information Systems Engineering. Resistencia Regional Faculty. National Technological University.
[2]Department of Informatics. Faculty of Engineering. Lomas de Zamora National University.

Abstract. In recent years, various methods to automate the quality control in software products have been developed. Nevertheless, there are few applications orientated web application assessment. Thus, this article describes the main features of a tool designed to assess quality evaluation of web software, called QUCO2, and implemented based on research work that was performed for the development a framework of web quality assessment.

Keywords: Quality, Web Quality Assessment, Quality Model.

1 Introduction

Quality assessment in products or services of any kind has been an everyday and increasingly important thing, since it became a differentiating factor when you choose to acquire a service or a good. The concept of quality has different definitions, but the widely accepted one is established by ISO 9000 [1], a norm which defines quality as "*the degree in which a set of inherent characteristics fulfills requirements*".

Software, as a product, should also be part of a quality assessment process, and thus, to determine the degree of satisfaction to the requirements and needs of the user. Web applications are a special type of software. They have characteristics that differentiate them from traditional systems such as: the size and complexity of the applications, the multidisciplinary nature of the development team, the hasty rate of the project delivery, among others. The above mentioned features bring with them the concept that the existing processes, models and metrics to evaluate the quality have to be adapted to consider the changes imposed by new technologies [2].

This paper presents the technical characteristics of a technological tool, QUCO2, in which the implementation aim is to automate the quality assessment of web applications from the point of view of the user. Its development is part of interinstitutional research project, "Models and Metrics for the Evaluation of Software Quality", that teacher researchers from the Faculties of UTN Regional Resistencia and Exact and Natural Sciences, and Surveying from UNNE carry it forward under the guidance of a researcher at the National University of Lomas de Zamora .

Section 2 presents the framework, which is still in development and is part of the aforementioned Project, in addition of the tool that is detailed in this publication, it

also comprises a quality model particularly oriented to web applications. Section 3 describes the technical design of the project, i.e., QUCO2 tool architecture and the functions they perform. Section 4 shows some of the results obtained with QUCO2 and a comparison of this with other existing tool. This comparison is summarized in a comparative table showing the differences between them and the advantages and disadvantages of each.

Finally, the last Section includes conclusions and future works that could be done to extend the functionality of the tool in development.

2 WQF: Framework for Quality Assessment in Web Applications

As it is mentioned in the previous section, QUCO2 is one component of WQF, a framework that allows us to evaluate the quality from the point of view of the product, particularly oriented towards web applications.

The framework is the result of research work of various models of quality, standards and norms such as ISO 14598 and ISO 9126. At the moment, these models are part of ISO 25000 standard that defines the way you should assess the quality of software products and the quality model to be followed [3][4][5][6][7]. Aspects to be evaluated from the quality software are grouped in Features, which in turn are shaped by metrics that are methods and scales for measurements and also, they are the result of mathematical relationships between parameters and specific attributes of the measurements. Thus, and as a result of research work, it has been developed a model of quality oriented to web applications [8]. However, as it is necessary to integrate the quality model and the results of the evaluation, it has also been designed a framework for this purpose.

This framework is called WQF, which to manage quality elements, includes a quality model (WQM) and a software tool (QUCO2), and it is developed based on this model.

2.1WQM: Quality Model for Web Software

In the first instance, and according to the scope of this research, the proposed quality model is made taking into account the following metrics [9]:

- Metric 1 - Usability: It is regarded as the degree of effectiveness, efficiency and satisfaction by which specified users can achieve specific objectives, in contexts of specific usage to use a product. The criteria to be evaluated are:
  a) Learning Facility: refers to the need to minimize the time required with respect to the learning curve Software.
  b) Consistency: a system is consistent if all the mechanisms remain the same circuit regardless of time.
  c) Recoverability: This aspect refers to the facility to correct an action once the user acknowledges an error in the operation.

d) Retention time: measures the ease, from the point of view of the user, to remember the operation of the system even though considerable time has elapsed since the last time it was used.
e) Flexibility: assesses the potential for Exchange of information between the user and the system.

- Metric 2: Reliability: This metric is related to the ability of software to maintain its level of performance under stated conditions for a period of time. The criteria to be evaluate are:
  a) Frequency and severity of failures: it measures how often failures occur in the system, if it occurs, and the ability of software to maintain the specified level of performance.
  b) Accuracy of the outputs: measure that indicates the approximation of the output obtained from the output achieved by the software.
  c) Failover capability: This measure includes processes required to detect and recover from abnormal situations. It should provide a minimum expiration time, after which you must apply new response.
  d) Reliability: measures the occurrence of unauthorized access to private information.
- Metric 3 - Functionality: This metric allows us to check the relationship between the functions of applications, the expected results and the real results. The quality criteria to be evaluated are:
  a) Adequacy: attributes that determine whether the set of functions are appropriate for the specified tasks.
  b) Safety: attributes that measure the ability to prevent unauthorized access, whether accidental or deliberate, both programs as data.
  c) Compliance: attributes that make software adheres to standards related to the application, and conventions or legal regulations.
  d) Reliability: No enabling of unauthorized access to private information.

Each metric proposed associates a weight with each feature (e.g. No Apply, Apply, Apply Heavily), evaluated on a scale of measurement (e.g. Wrong, Regular, Good, Very Good). The derived of the general formula to calculate the overall quality level we get the Framework is:

$$NO = \sum_{i=1}^{n}(VC.PC) / \sum_{i=1}^{n}(PC) \quad (1)$$

where NO is level obtained, VC is the calculated value for the metric i and PC is the weight of the i feature. Summations are perfomed based on all the components selected for the evaluations. Basically it is an average between the values obtained for each component influenced by the weight of that component in the overall study [9].

Thus, with this model it is possible to obtain a certain quality level of web applications from the evaluation metrics included in the WQM model.

3 QUCO2: Description and Technical Features

In order to manage the elements of the quality model and analyze the results of the evaluations, it is necessary an application development that permits the automation of these tasks. So, we worked on getting QUCO2, an application that enables to do web software evaluations, by different users, and offers information about the resulting quality value.

This section briefly describes each of the components used in the development of the application, providing a detailed look at the features in which each of them play in the execution of the application.

3.1 Application Design

QUCO2 development was posed as a web plugin, to facilitate the use of tool to the user, because it fits any browser and navigating different sites without interruptions in the execution of the application. Hence the user may, simultaneously, to review comprehensively the Web application to assess and record the required information. Thus, QUCO2 appears as a small container for small reusable components that enables to evaluate a particular feature. For each possible evaluated feature, it is defined a scale associated therewith, with the possibility of defining in turn a relative weight according to the project. Having the characteristics defined, scale and relative weight, one component is created to record and collect data for quality analysis. This means that you will get generic information and consolidated on product quality evaluation. After the evaluation process, the tool will provide the user with a quality level value obtained for the product in question.

In addition, the software will supply different user roles: *Developers*, who are responsible for the registration of information relating to the development process and the design of self-assessment to monitor the evolution of product quality; *Quality Assessors*, who have the responsibility to verify that the information recorded by the developers is correct and the quality assessments from their perspective, and *Customers*, who will display the information generated by the system and make decisions based on it.

From the functional point of view, the tool was developed under the client server philosophy in three layers. In this case, the user interacts with the web application through the browser. As a result of user activity, requests are sent to the server, which hosts the application and normally makes use of a database that stores all the information related to it. The server processes the request and returns the response to the browser that the user presents. Therefore, the system is divided into three components: the browser, which presents the user interface; the application, which is responsible for performing the necessary operations according to the actions made by this, and the database where the information related the application turns persistent [10]. This distribution is known as the model or architecture of three-tier, and is shown in Figure 1

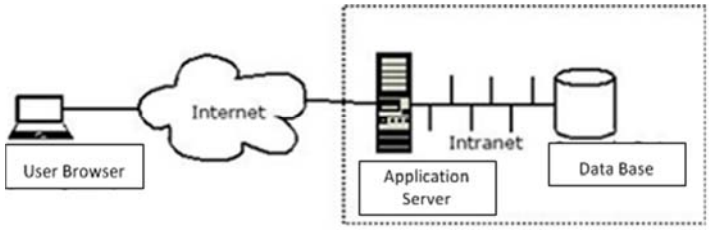


Fig. 1. Three-tiers architecture.

Applications developed under the three-tier architecture can be divided into the following levels:

1. **Presentation Level:** it is in charge of generating the user interface depending on the actions carried out by itself.

2. **Business Level**: it contains all the logic modeling business processes and is where all the processing necessary is done, to meet user requests.

3. **Data Management Level**: it is charged to make persistent all the information, supplies and stores information for the level of business. The following is a presentation on how integrated and implemented are these layers in the front-end and back-end QUCO2,

3.2 Front-End

In client-server applications, the client is the process that allows the user to formulate the requirements and pass them to the server. It is also known by the name of *front-end*.

In the case of QUCO2, particularly, the functions performed in the front-end are:

- Manage the User Interface: necessary adjustments are made to the interface application, the components are redesigned and the functionality is modified as new requirements arise.
- Interact with the user: that is, the communication process of the Evaluator with QUCO2.
- Process the application logic and making local validations: Each time a user enters the system, your profile will be check to enable the functions which are applied.
- Generate requirements to the database: Both when the user log as when issuing the request to save the assessment, requirements will be generated to the database
- Receive and display results from the server: the screen will show the user a variety of information of relevance, for example the quality level obtained as a final result.

  The design, in graphical form, the front-end of QUCO2, it is shown in Figure 2:

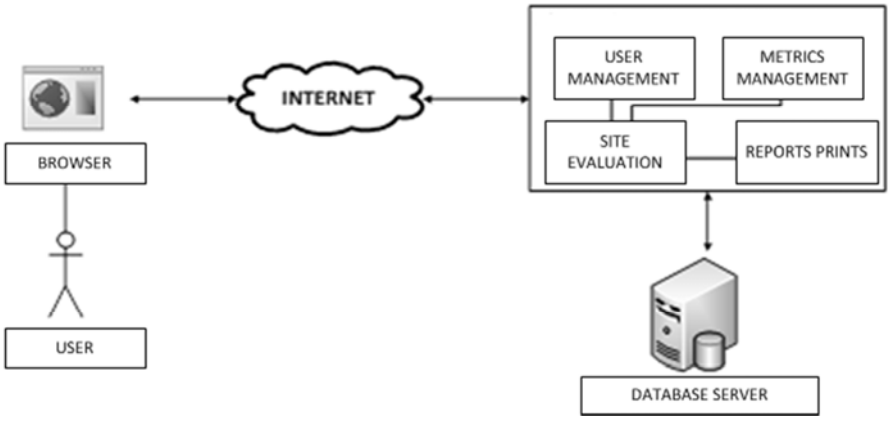


Fig. 2. Design Diagram of front-end..

To implement all required functions, before starting the development process, integral modules were designed where the functional elements of the applications will be distributed:

- User Management: It allow us to record information from different users that will interact with the software. Not only store data access account, but also the user profile according to the roles described in the previous section.
- Metrics Management: From here you create components using Quality Model. In this case WQM involves creating metrics and their relation with features that will have associated a relative weight. Then, this information is used to calculate the level of quality obtained in the evaluation of a product.
- Reports Issuance: This component is still in development, and allows the production of various statistical reports to expand knowledge about the current use of the tool. This functionality will be the most important for the Customers; then based on generated reports, decisions may be taken according to the observed results.
- Web Quality Assessment: From here, Assessors will may record the test they take, from their perspectives, from any web product. Also it will allow you to get the quality level obtained as a final result of the evaluation process.

The first two modules will only be accessed by the Developers, who are responsible for configuring the application for later use.

All these features are part of the general functionalities of the interface that is provided to users. Graphical and functional implementation of the application is carried out using v2.0.002 Twitter Bootstrap technology, consisting of a collection of free software tools for creating websites and web applications. It contains design templates based on HTML and CSS with typography, forms, buttons, graphics, navigation bars and other interface components, and optional extensions of JavaScript.Bootstrap that was developed by Mark Otto and JacbodThornton Twitter, as a framework to promote consistency through internal tools [11]. It was decided to use this technology to QUCO2, because it lets build compatible applications with

most web browsers. Also since version 2.0, it supports *sensitive designs*, i.e., that the graphical interface of the resulting product is dynamically adjusted, taking into account the characteristics of the device used for the execution (PC, tablets, smart phones, etc.) Finally, the main reason for the choice of Twitter Bootstrap had to do with CSS enhancements offered by a number of utilities javascript that facilitate user interaction. This technology need to use, in turn, the project Less [12] which allows a continuous cycle of development and improvement of CSS using advanced features. As an extension of CSS, LESS includes variables, mixes reusable code snippets, simple mathematical operations, nesting and even color functions. The combination of Bootstrap and Less user interface provides a pleasing and supplies a "responsive web design" that allows you its use on devices with limited display characteristics (e.g. smartphones) without redesigning the interface.

Each time the user performs an evaluation of a particular website using the plugin, and generates action to save the results of the transaction, the request is received by the Web server, and the information generated is stored by the user on Server database, recording the quality final value obtained..

3.2 Back-End

The functions performed in the back-end or server process, is basically provided in the implementation of business logic and storage of required information.

To develop the back-end system, which is plotted in Figure 3, we used the Symfony web framework [13], based on PHP, using as Propel project mapper [14]. This combination lets the deployment of changes of the system, in a very quickly and efficiently way, supporting schema modifications of the database, and the implementation of new business rules or even complete redesigns interface in a very transparent way, i.e., without affecting the interaction with the user.

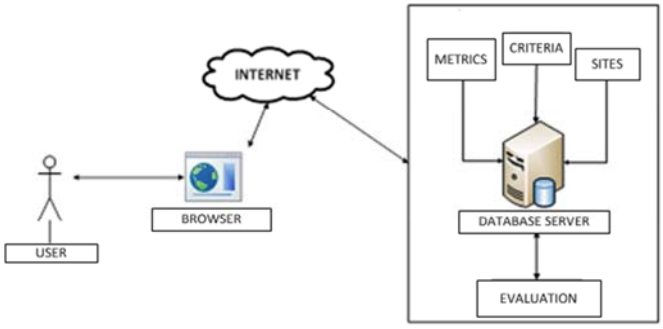


Fig. 3. Overall Design Diagram of the Back-End

All the information that is recorded and obtained by evaluation process, it will be stored in a MySQL database, the choice was based on the following issues:

- MySql is Open Source software, licensed under the GPL.
- The technology provides speed and security for operations, it is essential in database applications on the web.
- Easy installation and configuration of the database on the web server.
  To create the configuration of the different options for each metric, the system uses the interchange format JSON data [15], which has much support across platforms and it is easy to understand and learning in case of need for new metrics.

In turn the whole project is published on a Github repository, enabling up projects using the version control system Git, under the MIT license [16] which enables sharing with the development community under the form Open Source.

4 Results Obtained. Comparison with other tools.

Currently a variety of tools aimed at evaluating the usability of software products. However, as a result of this investigation, it was verified that none of the technologies aimed at the evaluation of web applications are defined according to criteria of any existing quality model, nor returned to the user a final result of the quality level obtained. Among all the existing tools in the market, we have chosen two of them for comparison with QUCO2. The first is Alexa [17], which returns information about the visits to a domain, while the position calculated by global positioning ranking and regional level taking into account the popularity and visitor traffic. The other tool is developed by Silktide, called Nibbler [18], which has the distinction of providing disaggregated score in "accessibility", "technology", "user experience" and "marketing" of the web, and indicates how to improve in every aspect.

To evaluate the partial results obtained from the use of QUCO2, and to compare with those generated with the other two tools, we have defined a context of use, taking into account the commercial field and state. So, three general interest websites are chosen for each environment: a) an e-commerce site, b) an University portal, and c) an Governmental site, being one of the most widely used government sites among Argentines. The evaluation team was formed with 40 people. Regularly, all of them expressed assiduity in the use of the mentioned websites and the knowledge regarding the same, in terms of functionality, interface, etc.

By analyzing the results in the case of the presented study, it was demonstrated the correct operation of the tool in the evaluation process. The use of the plugin did not present difficulties in their learning curve, according with the feedback obtained from the participants.

Once the assessment finished with QUCO2, we calculated a weighted average of the values obtained for each proposed site. Based on a balancing of the weights of the metric considered, it was observed that the maximum satisfaction value occupies the range between 20 and 22, the average quality value corresponds to the range between 14 and 20, the value of average quality ranks between 8 and 14, and minimum value recorded level below 7. Given these levels, the results of evaluations are presented below this:

a) In the case of www.mercadolibre.com site, using information from QUCO2, there was a quality average value obtained that was 20. It is also evaluated as an *excellent* quality site according to the quality model described in section 2 Using Nibbler, the obtained value was 5.1, and Alexa positions is in globally ranked number of 2257.
b) For www.anses.gob.ar site, QUCO2 yields an average value of 17, then valuing quality as a *very good* site. Using Nibbler, the value obtained is 8.4, and Alexa position is globally in the post 13794.
c) In the case of the UTN portal, www.utn.edu.ar, the tool obtains an average value of quality 14, i.e. a *regular* quality site. Nibbler gets a value 3.6 and Alexa position is globally in the post 33198.

5 Conclusions and Future Works

As a result of partial validation and considering the comparison of the evaluation results obtained using QUCO2, and against the values obtained using the other two tools already available on the market, there is similarity in the levels of quality for websites of the case studied. However, unlike other applications, it was found that AUCO2 represents an interrogative tool that gets its final evaluation value considering a set of metrics defined in the model WQM of WQF, and not focusing on quality aspects or isolated quality criteria. It also stresses that being an open-source implementation facilitates their implementation and use in any environment.

As future works, it is intended to continue the development of the framework, including first, the missing metrics (Maintainability, Safety, Availability and Scalability) of the quality model and to the software tool, according to the needs presented by web applications. QUCO2 also aims to bring all kinds of software, according to defined quality in different models.

6 References

1. ISO. “Systems of Quality Management – Concepts and Vocabulary”. Internacional Norm ISO 9000,2000.
2. Abrahao, Silvia; Pastor, Oscar; Olsina, Luis; Fons, Joans. “A method for measuring the functional size and assess the quality of websites”. Group I+D in Software Engineering (GIDIS). Faculty of Engineering, UNLPalm. La Pampa, Argentina.
3. ISO, “ISO/IEC 14598 – Software Product Evaluation” (2001)
4. ISO, “ISO/IEC 9126-1 – Software engineering–Product quality – Part 1: Quality Model” (2001)
5. ISO, “ISO/IEC 9126-2 – Software engineering– Product quality – Part 2: External Metrics” (2003)
6. ISO, “ISO/IEC 9126-3 – Software engineering– Product quality – Part 3: Internal Metrics” (2003)
7. ISO, “ISO/IEC 9126-4 – Software engineering– Product quality – Part 4: Quality in Use Metrics” (2003)

8. Martínez, Nelson Enrique León; Chacon Pinto, Nelson. “Computational tool for evaluation of software product quality framed in research”. Pereira Technological University.
9. Pinto, Noelia; Tortosa, Nicolás; Acuña, César; Cuenca Pletsch, Lilian; Estayno, Marcelo. “Evaluation of Web Application Quality assisted by technological tools”. WICC 2013. ISBN 978-987-28179-6-1
10. Hernández, Edgar; Martínez, Luis. “Client / Server”. Technological Innovation Club. San José, Costa Rica.
11. Proyecto Twitter Bootstrap V2, `http://twitter.github.com/bootstrap`
12. Proyecto Less, http://lesscss.org
13. Symfony, Framework de Desarrollo Web, http://www.symfony-project.org/
14. Proyecto Propel, http://www.propelorm.org
15. Proyecto JSON, http://json.org
16. GitHub, http://www.github.com
17. Software Alexa, http://www.alexa.com
18. Software Nibbler, http://nibbler.silktide.com/

Análisis de la información presente en foros de discusión técnicos

Nadina Martínez, Gabriela N. Aranda, Mauro Sagripanti, Pamela Faraci, Alejandra Cechich

Grupo GIISCo, Facultad de Informática, Universidad Nacional del Comahue
Buenos Aires 1400 (8300) Neuquén, Argentina
{nadina.martinez|gabriela.aranda}@fi.uncoma.edu.ar

Resumen Los foros de discusión se han convertido en la herramienta colaborativa más utilizada por los practicantes informáticos para realizar preguntas y recibir propuestas de otros técnicos para solucionar o mejorar problemas técnicos particulares. Con el objetivo de construir un navegador especializado en encontrar dichas soluciones, este artículo introduce un modelo de la información contenida en foros de discusión técnicos y presenta los resultados preliminares de una encuesta realizada a usuarios de dichos foros, enfocada en la percepción de la adecuación de los hilos de discusión a un problema y la correctitud de las soluciones propuestas.

1. Introducción

La incorporación de la tecnología personal ha traído como consecuencia que más personas estén dispuestas a hacer uso de ella, obteniendo directamente beneficios tangibles [1]. Dentro de la comunicación virtual, las herramientas colaborativas como redes sociales, wikis, blogs y foros de discusión permiten el intercambio de información de forma rápida y eficaz, acortando las distancias entre las personas que tienen el conocimiento y aquellas que lo necesitan. Estas herramientas colaborativas no son utilizadas en forma personal solamente, sino también en entornos laborales e incluso académicos. Las herramientas pueden clasificarse en sincrónicas y asíncronas (según se requiera, o no, que las personas estén conectadas al mismo tiempo) [2]. Aunque en los ámbitos laborales y académicos pueden utilizarse ambos tipos de herramientas, se suelen aprovechar las asíncronas para diseminar conocimiento, ya que permiten que la información esté disponible aún cuando las personas que tienen el conocimiento no estén alcanzables.

Aún cuando hay varias herramientas colaborativas asíncronas disponibles en la web, existen algunas diferencias entre ellas. Por ejemplo, los blogs tienen una naturaleza no interactiva, donde una persona que es el dueño (autor del blog), escribe una bitácora o diario en línea, permitiendo que los visitantes participen agregando comentarios, no habiendo comunicación entre los participantes. Otra herramienta asíncrona colaborativa son las Wikis, que son páginas en las que

un autor publica algún tipo de información, a partir de ese momento los otros usuarios que acceden a dicha Wiki pueden modificarla (con la autorización del autor y dependiendo de la privacidad que ofrece el sitio). Por el contrario, los foros de discusión son canales de comunicación cuya finalidad suele ser intercambiar información, experiencias y conocimiento entre sus usuarios. En general son informales y dependen de un moderador (persona que mantiene el orden y naturaleza del foro). Una característica particular de los foros es que cualquier usuario puede comenzar un hilo de discusión quedando establecido un intercambio de información sobre un tema, permaneciendo de esta forma disponible para cualquier lector, por lo que un foro constituye una base de conocimiento al alcance del público en general. Particularmente, se han elegido los foros de discusión como base de este trabajo, dada su capacidad de representar problemas de los usuarios en general (no solo de los dueños de blogs o grupos de colaboradores en wikis), y dado que permite ver todos los comentarios y obtener conclusiones a partir de ellos (a diferencia de las wikis que esta información se mantiene oculta al público en general).

En la actualidad, cuando una persona tiene un problema técnico, ingresa una serie de palabras en algún buscador multipropósito, y éste devuelve una lista de enlaces a páginas Web de distinto formato (manuales, páginas de instituciones técnicas, blogs personales, foros de discusión, etc.) que contienen esas palabras. Luego la persona interesada va observando cada elemento de la lista, y debe visualizar el contenido de cada página para determinar si éste le sirve o no. Esta lista de elementos está ordenada de acuerdo a políticas del buscador, la cual puede no ser precisamente el orden de importancia que el usuario necesitaría. El objetivo futuro al que apunta nuestro trabajo es construir un navegador especializado en problemas técnicos que, a partir de un conjunto de palabras clave que representan la búsqueda inicial, retorne una lista ordenada de soluciones candidatas. Dichas soluciones se obtendrán a partir del análisis previo de varios hilos en foros de discusión técnicos. El orden otorgado a las soluciones candidatas será determinado por medio de un proceso de evaluación de calidad de la información. Con este objetivo en mente, nuestro trabajo actual se ha enfocado en analizar cómo los humanos acostumbran a buscar información en un hilo de un foro de discusión. Para ello, el resto del artículo está organizado de la siguiente manera: primero se introduce un modelo conceptual que representa la información contenida en los foros de discusión. Posteriormente se presenta el cuestionario preparado para recolectar conocimiento tácito de usuarios habituales de foros de discusión técnicos y la manera que ellos seleccionan qué soluciones probar. A continuación se presentan algunos resultados preliminares de la aplicación de dicho cuestionario. Por último se presentan las conclusiones y líneas de trabajo futuro.

2. Modelo conceptual para foros de discusión técnicos

Para definir el punto de vista de nuestro trabajo, se han tomado como base la clasificación de las necesidades de los actores relacionados con un sitio Web [3] y

la clasificación de los usuarios de foros de discusión de Roquet [4], destacando al usuario administrador como aquel usuario con mas privilegios que tiene el control total el foro; que determina quiénes serán los usuarios con roles de moderadores. También son importantes los usuarios moderadores ya que no sólo monitorizan las conversaciones sino aseguran que se cumplan las reglas de convivencia entre el resto de los usuarios. Los usuarios que publican, preguntan y contestan en los foros son los participantes, y por último se llaman participantes externos las personas que sólo pueden leer las conversaciones establecidas en los foros.

Con el objetivo de reutilizar el conocimiento contenido en las conversaciones entre usuarios participantes de una comunidad virtual como es un foro de discusión sobre temas técnicos, la primera instancia es definir un modelo de calidad para la información contenida en dicho tipo de foros. En este sentido, es pertinente que el modelo se plantee considerando sólo el punto de vista del *usuario externo,* es decir, enfocándose en la calidad desde el punto de vista de la información y no de la funcionalidad que el sitio pueda o necesite proveer para el resto de los tipos de usuario.

2.1. Esquema conceptual

Con el fin de establecer un marco teórico para el estudio de la información contenida en los foros de discusión, se realizó una revisión formal de 36 hilos de discusión reales en 6 foros distintos en idioma español e inglés. El resumen de dicho análisis se presentó en [5].

En base a dicho análisis se propuso un primer modelo conceptual de la información disponible en un foro de discusión desde el punto de vista del usuario externo, identificándose las entidades más importantes y sus atributos. A continuación se presenta una actualización de dicho modelo, al cuál se ha agregado el tipo de fragmento de mensaje *figura*. Dicha actualización del modelo surge a partir de la extensión de la revisión, al abarcar foros de discusión que permiten mostrar capturas de pantalla y otros tipos de imágenes en formato gráfico (los cuales no habían sido cubiertos en la primera revisión). Otra mejora al modelo se ha realizado en la definición de los atributos de los usuarios, para los cuales se ha diferenciado su *experiencia*, indicando el reconocimiento de la comunidad del foro respecto al nivel de pericia de dicha persona en el tema de discusión (generalmente expresado en una escala como novato, experto, gurú, etc), y por otro lado la *reputación*, que puede o no estar relacionado a la experiencia, y se suele representar en los foros como agradecimientos, pulgares arriba o abajo, indicadores de "me gusta", etc. Ambos atributos suelen ser expresados verbalmente o con imágenes tipo icono según el foro estudiado.

En la Figura 1 se presenta el modelo conceptual actualizado de acuerdo a los considerandos explicados previamente.

El modelo conceptual puede resumirse de la siguiente manera: un foro de discusión técnico (*foro*) contiene varios hilos de discusión (*hilo*). Cada *hilo* se genera cuando un usuario participante de la comunidad (*usuario*) crea un nuevo tema de debate que surge generalmente a partir de una inquietud personal. Cada *hilo* se identifica por un *título*, que está generalmente relacionado con la

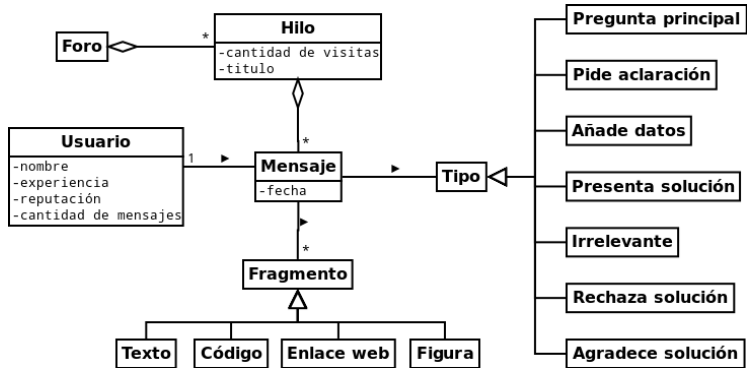


Figura 1. Diagrama de la información contenida en un foro de discusión técnico

pregunta principal, realizada por el usuario que inicia el hilo o tema de debate (esto suele ser un requisito mencionado en las reglas del foro y controlado por los moderadores de los foros). Si bien es cierto que no está presente en todos los foros, suele encontrarse disponible la información relacionada con la *cantidad de visitas* realizadas al *hilo*, es decir la cantidad de veces que la página fue accedida o visitada por un usuario participante o externo.

La estructura del *hilo* está formada por una serie de aportes. Cada aporte, llamado post o mensaje (*mensaje*), es realizado por un *usuario* participante en una *fecha* en particular. A fin de poder analizar el contenido de cada *mensaje*, se ha determinado que un *mensaje* consta de uno o más *fragmentos*, donde cada fragmento puede tratarse de lenguaje natural (*texto*), código que puede ser ejecutado en un sistema operativo o compilado en un lenguaje de programación (*código*), un enlace a una página Web donde una pregunta similar con posibles soluciones afines han sido propuestas (*enlace web*). También puede ser una *figura* (contenido en formato gráfico de tipo jpg, png, etc) que suele utilizarse para incluir esquemas, diagramas, capturas de pantalla, etc.

En base al análisis semántico de los fragmentos que componen el *mensaje*, se definió que existe un mensaje correspondiente a la pregunta principal y otros que cierran o completan la conversación: uno o más mensajes que proponen soluciones, que rechazan alguna solución, y otros que cierran la pregunta de manera positiva (agradeciendo alguna solución). Además de estos mensajes principales, se han reconocido los siguientes: mensajes de pedido de aclaración, de agregado de nuevos datos (que ayudan a los que responden a situarse en el problema), y mensajes irrelevantes (que a veces suelen ser eliminados por el moderador pero otras veces son parte del hilo de la conversación).

Respecto a los usuarios, para cada *mensaje* se puede saber el *usuario* que lo escribió, del cual se conoce su nickname o nombre dentro de la comunidad (*nombre*). Si bien no es un dato presente en todos los foros de discusión, habitualmente se cuenta con más información sobre el usuario como su *experiencia* y *reputación* (que fueron explicados anteriormente) y la *cantidad de mensajes* que ha emitido en la historia de su participación en la comunidad.

3. Encuesta a usuarios de foros de discusión técnicos

Con el objetivo de definir criterios para estimar la calidad de la información contenida en los foros de discusión técnico, se han definido dos características principales. La primera es la *pertinencia* de un hilo de discusión, es decir, el grado de proximidad entre el problema discutido en un hilo de discusión y el problema original definido por un usuario (expresado a partir de una cadena de búsqueda determinada). Y la segunda característica importante es cuán adecuada o *correcta* es una solución para el caso particular del usuario interesado.

En base a estas dos características se plantearon las siguientes preguntas principales para la investigación:

- *¿Cómo determinan los usuarios de los foros de discusión técnicos qué hilos leer (y cuales no)?*
- *¿Cómo seleccionan los usuarios de foros de discusión técnicos las soluciones a probar?*

Para resolver dicha pregunta principal, se propusieron las siguientes sub-preguntas:

- ¿Qué información consideran importante los usuarios para determinar la pertinencia de un hilo de discusión?
- ¿Con qué frecuencia consideran que un ítem de información es importante para determinar la pertinencia de un hilo de discusión?
- ¿Qué información consideran importante los usuarios para determinar la correctitud de una solución propuesta en un hilo de discusión?
- ¿Con qué frecuencia consideran que un ítem de información es importante para determinar la correctitud de una solución propuesta en un hilo de discusión?

3.1. Definición y aplicación del cuestionario

Para responder las preguntas planteadas, se definió un cuestionario destinado a usuarios habituales de foros de discusión de tipo técnico. La estructura del cuestionario fue la siguiente:

Primera Sección: además del nombre y rango de edad de los encuestados, se incluyeron las siguientes preguntas:

- ¿Qué rol o roles relacionados a la informática cumple habitualmente?
- ¿Con qué frecuencia accede a foros de discusión técnicos?
- ¿A qué temáticas se refieren los foros de discusión técnicos que visita?

Segunda Sección: en esta etapa se solicitó elegir una opción en la escala [Siempre, Casi siempre, A veces, Casi nunca, Nunca], para once afirmaciones que marcan la importancia de los ítems de información definidos en el modelo conceptual. Por cuestión de espacio, a continuación se exponen como ejemplo las dos primeras:

- Si el título del hilo tiene todas las palabras claves ingresadas, alcanza para saber si el tema en cuestión está relacionado con mi búsqueda.
- Si la pregunta principal (primer post) está relacionada en parte con lo que estoy buscando continúo leyendo el resto del hilo.

Tercera Sección: en esta sección se solicitó que los encuestados seleccionen uno o más ítems de información definidos en el modelo conceptual (título, mensaje principal, fecha del mensaje, etc), relacionados a las siguientes consignas:

- Para estimar si un hilo de un foro de discusión está relacionado con mi problema (es pertinente), observo...
- Suponiendo que se trata de un hilo que es pertinente para su problema, para estimar si una solución propuesta es correcta (correctitud), la información que observo es...

Finalmente, en la última sección se dejó espacio disponible para que los encuestados incluyeran comentarios o sugerencias de distinto tipo.

El cuestionario fue implementado mediante un formulario (*form*) en la plataforma de Google Drive[1]. Además, en el sitio web del proyecto[2] se publicó una página donde se explica el objetivo de la encuesta y se describen las secciones y pautas establecidas para el cuestionario. El enlace a dicha página se envió por correo electrónico a un conjunto de 40 personas que cumplen distintos roles relacionados a la informática en el ámbito de la Universidad Nacional del Comahue (docentes, estudiantes y graduados desempeñándose en el ámbito laboral local). Al momento de la redacción de este artículo se cuenta con 24 respuestas, cuyos resultados serán presentados en la siguiente sección.

3.2. Resultados preliminares

Perfil de los encuestados. A partir del análisis de la primera sección del cuestionario, se puede definir si el perfil de los encuestados abarca un amplio registro de tipos de usuarios de foros de discusión.

En primer lugar se observa que la mayoría de los encuestados son usuarios habituales de foros de discusión técnicos. De acuerdo a la Figura 2, el 75 % de ellos accede varias veces a la semana (58 %) e incluso algunos diariamente (17 %). Esto es importante para este estudio, dado que permite confiar en el conocimiento previo de los encuestados al momento de responder las preguntas planteadas.

Respecto a los roles que estos encuestados cumplen, puede observarse en la Figura 3 que se han visto representados todos los roles excepto el de tester, lo cual deberá ser tenido en cuenta al extender la muestra de usuarios en el futuro. Los porcentajes más altos corresponden a los roles de programador (20 %) y docente (15 %). Sólo dos encuestados informaron un rol distinto a la lista inicial: uno es

[1] https://drive.google.com/

[2] http://forumadvisor.wordpress.com/encuesta-calidad-de-la-informacion-en-foros-de-discusion-tecnicos/

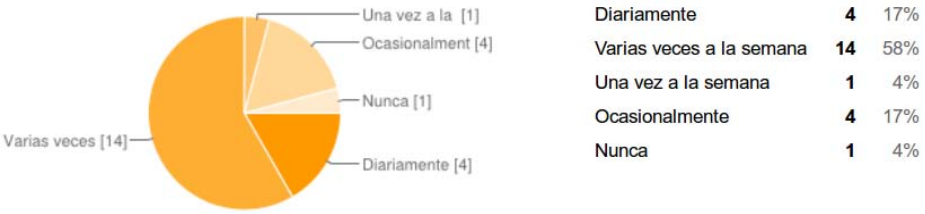


Figura 2. Frecuencia de acceso a foros de discusión

"Administrador de sistemas y redes" y el otro "Administrador de infraestructura de sistemas". Las tareas de dichos roles deberán ser analizadas y considerar la inclusión de ambos roles en una nueva aplicación del cuestionario.

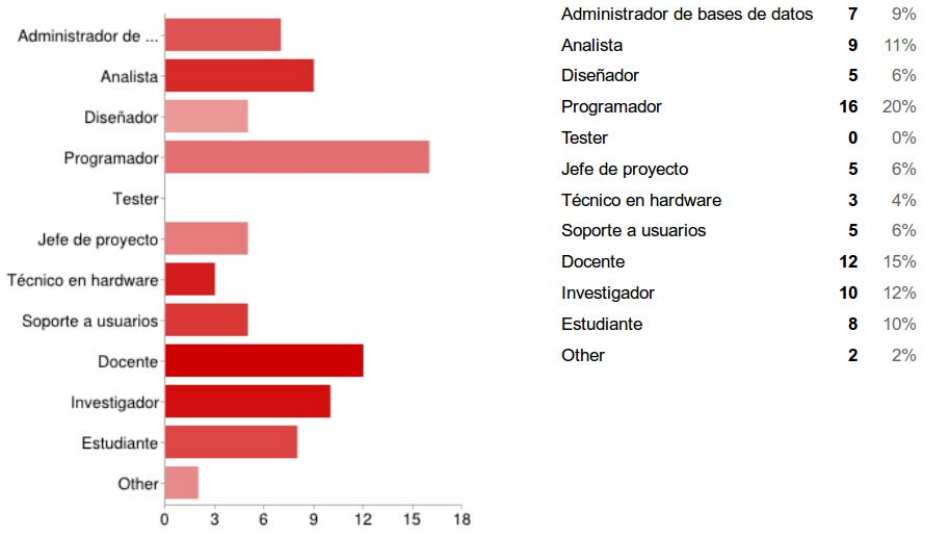


Figura 3. Roles de los encuestados relacionados a la informática

Analizados los tipos de foros visitados, puede observarse en la Figura 4 que todos los tipos propuestos han sido abarcados por los encuestados. Los porcentajes más altos corresponden a los foros sobre lenguajes de programación (31 %) y herramientas de software específicas (24 %). De las restantes, el rubro de foros sobre desarrollo de aplicaciones Web es el más utilizado (14 %). Otra vez, dos tipos de foro fueron agregados por los encuestados en la opción "Otros" y corresponden a los mismos usuarios que agregaron un nuevo rol a la lista. Los tipos de foros agregados son "Administración de sistemas" y "Active Directory". Dichos tipos de foros deberán ser analizados para considerar su inclusión en una nueva aplicación del cuestionario. Además, en el futuro planeamos extender el análisis de la información recolectada para comprobar estadísticamente la correlación entre los roles de los encuestados y los tipos de foros visitados por ellos.

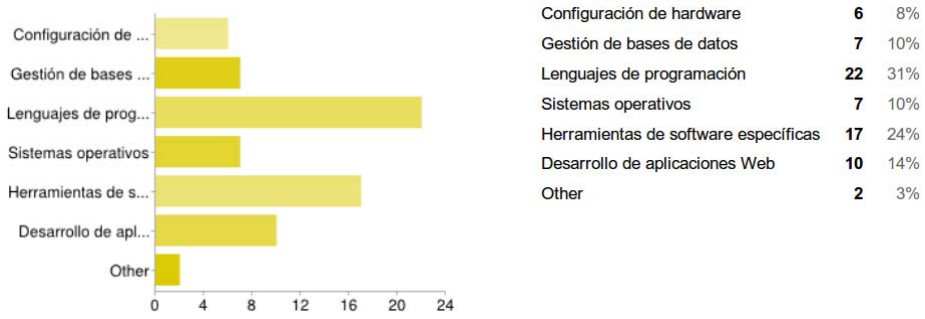


Figura 4. Temática de los foros visitados por los encuestados

Análisis de la pertinencia de la consulta. A continuación se analiza la información de la primera pregunta de la tercera sección del cuestionario.

Esta pregunta pedía a los encuestados que seleccionaran aquellos ítems (uno o más) que chequean para definir si un hilo de discusión está relacionado con su problema particular. Los resultados obtenidos se muestran en la Figura 5.

Figura 5. Información analizada para definir la pertinencia

Los ítems más seleccionados son el título (19 %), la pregunta principal (21 %) y los mensajes que proponen las soluciones (17 %). Al contrario de nuestra expectativa previa, 3 encuestados no seleccionaron la pregunta principal como lo más importante para definir si el hilo era pertinente o no, e incluso el título fue elegido por más personas (22) que la pregunta principal (21). Sin embargo, a partir de análisis de cada caso, debe destacarse que todos los encuestados seleccionaron el título o la pregunta principal como importantes para definir la pertinencia del hilo. Algo que también debe destacarse es el bajo porcentaje otorgado a la experiencia previa y la reputación de los usuarios que proponen la

solución, siendo apenas un 4 % o 5 %, mientras que los mensajes que agradecen una solución son utilizados como referencia más a menudo (9 %).

Discusión. En base al análisis de un conjunto preliminar de encuestas, se puede resumir que el perfil de los usuarios encuestados abarca un amplio espectro de roles relacionados a la informática. Ellos, a su vez, visitan habitualmente varios tipos de foros de discusión técnicos. Esta característica es importante para nuestro trabajo, dado que permitirá obtener una visión amplia del comportamiento de técnicos que buscan soluciones en foros de discusión. Respecto a los ítems de información que los encuestados indican importantes para identificar la pertinencia de un hilo en cuestión, es importante destacar que el título y la pregunta principal son los dos ítems más mencionados, y que, aunque la expectativa previa era que todos respondieran que la pregunta principal era siempre la más utilizada, 3 personas (4 %) no lo perciben así. Otro resultado interesante es que muy pocos encuestados (4 %) respondieron que consideran importante la experiencia o reputación de quienes proponen las soluciones. Estas tendencias deberán ser tenidas en cuenta al avanzar en la extensión del estudio.

4. Trabajos relacionados

La propuesta de Tigelaar et al [6] se enfoca en simplificar el contenido de los hilos de discusión extensos, resumiéndolos automáticamente con un prototipo de implementación basado en lenguaje natural. Este trabajo es un gran aporte para el análisis de los hilos de ejecución, pero no se enfoca en determinar si el conocimiento será de interés para la persona que lo consulta o no, como es el interés de nuestra propuesta.

En cuanto a reuso de conocimiento en foros de discusión, Chen et al [7] proponen un sistema recomendador para conocimiento desarrollado de manera colaborativa, analizando automáticamente los mensajes de un foro de discusión de un curso de Inteligencia Artificial para proponer mensajes con contenido similar, escritos por estudiantes de dictados anteriores del mismo curso. Otra propuesta existente es la de Helic y Scerbakov [8], que presenta un método de clasificación de los mensajes de un foro de discusión de acuerdo a una jerarquía de temas preestablecida. En primer lugar, nuestro enfoque se diferencia de las propuestas anteriores porque ambas están desarrolladas para dominios de aprendizaje colaborativo (e-learning), mientras que nuestro recomendador apunta a un dominio más amplio, que involucra usuarios con distinto conocimiento previo (background). Además, en dichos trabajos el foro utilizado es único, lo que permite asegurar que la información a analizar se encuentra en un formato estándar y que cualquier modificación puede ser prevista y gestionada a priori. Por el contrario, nuestra propuesta apunta a recolectar información de distintos foros, por lo tanto la heterogeneidad de formatos de la información a capturar y la posibilidad de cambios no programados es un desafío extra.

5. Conclusiones y trabajo futuro

En este trabajo se ha presentado una mejora del modelo conceptual para la información contenida en foros de discusión técnicos presentada en [5]. Luego, se ha introducido un cuestionario para usuarios habituales de foros de discusión técnicos y se han presentado los resultados preliminares sobre el perfil de los encuestados y el análisis de una pregunta del cuestionario, que considera la pertinencia de un hilo de discusión en relación con un problema particular. Como trabajo a futuro se planea avanzar en el análisis semántico de los mensajes en los foros, así como en establecer una serie de métricas e indicadores de calidad que sirvan para la detección automática de soluciones a problemas técnicos.

El objetivo a futuro es trabajar en el desarrollo de un buscador especializado en soluciones a problemas técnicos. Dicho buscador está previsto que mantenga una base de datos de las experiencias de los usuarios (después de seleccionar y aplicar las soluciones candidatas), como un mecanismo de mejora constante a partir de la retroalimentación realizada por los mismos usuarios.

Agradecimientos

Este trabajo está parcialmente soportado por el subproyecto "Reuso de conocimiento en foros de discusión técnicos", correspondiente al Programa de Investigación 04/F001 "Desarrollo orientado a reuso", de la Universidad Nacional del Comahue, y por el Proyecto PICT-2012-0045 "Mecanismos de soporte para grids híbridos orientados a servicios y técnicas de desarrollo de aplicaciones".

Referencias

1. S. Poltrock and J. Grudin, "CSCW, groupware and workflow: experiences, state of art, and future trends," in *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, (New York, NY, USA), pp. 120–121, ACM, 1999.
2. C. A. Ellis, S. J. Gibbs, and G. L. Rein, "Groupware: Some issues and experiences," *Communications of ACM*, vol. 34, no. 1, pp. 38–58, 1991.
3. L. Mich, M. Franch, and G. Cilione, "The 2QCV3Q quality model for the analysis of web site requirements," *Journal of Web Engineering*, vol. 2, pp. 105–127, Sept. 2003.
4. G. Roquet García, "Los foros de discusión en educación," *Siglo XXI: Perspectiva de la Educación desde América Latina*, no. 4, pp. 69–78, 1998.
5. G. Aranda, N. Martinez, P. Faraci, and A. Cechich, "Hacia un framework de evaluación de calidad de información en foros de discusión técnicos," in *ASSE 2013-Simposio Argentino de Ingeniería de Software, JAIIO 42Âº-Jornadas Argentinas de Informática*, (Córdoba, Argentina), p. a publicarse, SADIO, 2013.
6. A. S. Tigelaar, R. Op Den Akker, and D. Hiemstra, "Automatic summarisation of discussion fora," *Natural Language Engineering*, vol. 16, pp. 161–192, 4 2010.
7. W. Chen and R. Persen, "A recommender system for collaborative knowledge," in *2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, (Amsterdam, The Netherlands, The Netherlands), pp. 309–316, IOS Press, 2009.
8. D.Helic and N. Scerbakov, "Reusing discussion forums as learning resources in wbt systems," in *IASTED International Conference Computers and Advanced Technology in Education*, (Rhodes, Greece), pp. 223 – 228, 2003.

Knowledge Management in Distributed Software Development: A Systematic Review

Fernanda Tamy Ishii[1] , Gislaine Camila L. Leal[1], Edwin V. Cardoza Galdamez[1], Elisa Hatsue M. Huzita[1], Renato Balancieri[1] and Tânia Fátima C. Tait[1],

[1]State University of Maringá, Maringá, Paraná, Brasil
fernanda.tamyi@gmail.com,{gclleal, evcgaldamez, ehmhuzita, rbalancieri2}@uem.br, tait@din.uem.br

Abstract. Software development is characterized as a knowledge intensive activity. Particularly, Distributed Software Development (DSD) is an approach that demands more attention for coordination and communication among members of distributed team, due to regional, cultural and infrastructure differences. Knowledge has being, increasingly, seen as the most important strategic resource in organizations. So, the management of this knowledge is critical to organizational success. Knowledge Management (KM) is a set of processes directed at creating, capturing, storing, sharing, apply, and reuse of knowledge, which are useful to decision making. The purpose of this paper is to present a systematic review carried out to identify the processes, techniques, methods, practices and/or tools adopted for Knowledge Management in Distributed Software Development. With this systematic review some interesting points for research were identified.

Keywords: Knowledge Management; Distributed Software Development; DSD; Cooperative Development; Systematic Review.

1 Introduction

The Distributed Software Development (DSD) is an approach for software development that comes to meet of globalization need, such as: increase of productivity; quality improvement and costs reduction. It added to software development challenges concerned to cultural differences, geographic dispersion, coordination and control, communication and team spirit, which intensified some problems found during the project lifecycle [1], [2].

Software development is, by itself, characterized as a knowledge intensive activity, including for example the knowledge about processes, products, skills of different professionals involved. The interaction and information sharing among teams members are important parts of the process of knowledge construction in DSD context. Furthermore, knowledge has being seen as the most important strategic resource in organizations.

Knowledge Management postulates that organizations must deal knowledge as a factor of richness. It is a discipline that promotes in an integrated way, management and sharing of all information assets that a company owns. This information can be found in a database, source code, documents, procedures, as well as on people,

through their experiences and skills. So, when software development is considered, it is essential to maintain the knowledge available for fast and easy access for whose that needs it in the organization or in a supply chain.

In this sense, knowledge management has been increasingly used in companies to organize, strategically, the knowledge of employees and external knowledge, which are essential for business success. This paper describes a systematic review that was carried out in order to identify the processes, techniques, methods, practices and / or the tools adopted to promote knowledge management in distributed software development environment.

The text is structured into more three sections beyond this introduction. The second section describes the methodology adopted. The third section presents the results and discussion. Finally, the fourth section, presents the final considerations.

2 Methodology

A. A Systematic Review

Systematic review is a planned and structured process that aims at finding work and research related to a specific research question or concern. One of its applications is best known in the evaluation, interpretation and synthesis of available data on given technology, treatment or procedure, allowing the identification of problems and the formulation of a new scenario or solutions for this study [4].

B. Method adopted

The procedure applied to this review was adapted from the model presented in Kitchenham [4]. Firstly the research purpose was defined. Based on that, the research questions, which contain criteria for inclusion and exclusion of papers were defined. After that, the database to be used to carry out searches, the language, the publication year were also defined.

For data collection, were considered keywords related to issues of research and based on them, query sequence were formulated. All papers found were archived and cataloged with the help of a program called JabRef. After registration of the papers, began the pre-selection stage of them, which consisted of rapid reading of titles, keywords and abstracts. With this, was evaluated the adequacy of them regarding of search criteria established. After the pre-selection stage, was performed a complete reading of the papers approved in the first stage. Then the synthesis and interpretation of data were performed.

C. Research Goals

This review aimed to:

- Identify the processes, techniques, methods, practices and / or tools adopted for knowledge management in distributed software development.
- Observe the results of the practices of Knowledge Management in Distributed Software Development.

D. Research Questions

Considering the goals above mentioned, questions containing appropriated criteria for inclusion and/or exclusion of papers were elaborated. These are:

Question 1: What processes, techniques, methods, practices and / or tools are adopted to promote knowledge management in distributed software development?

Question 2: What are the benefits of the knowledge management in distributed software development?

Question 3: What kinds of difficulties, limitations and / or problems occur in the use of knowledge management for distributed software development?

E. Search Strategy

- Sources: Electronic indexed database (IEEE, ACM, Compendex and ScienceDirect).
- Language: English, once it is the internationally accepted language for scientific papers.
- Types of documents: conference proceedings, journal papers, books / book chapters, thesis and dissertation chapters and review reports.
- Year of publication: published papers in the period from 2000 to 2013.

G. Query strings

The keywords were determined considering the terms "distributed software development" and "knowledge management". The query string was create using logical operators AND and OR.

("Distributed Software Development" OR "Global Software Development" OR "Geographically Distributed Development" OR "Collaborative Development" OR "Distributed Development" OR "Distributed Software Project" OR "Global Software Engineering" OR "Globally Distributed Work" OR "Distributed Teams" OR "Global Software Teams" OR "Virtual Teams") AND ("Knowledge Management" OR "Knowledge Acquisition" OR "Knowledge Achievement" OR "Knowledge Retention" OR "Knowledge Application" OR "Knowledge Sharing" OR "Knowledge Use" OR "Knowledge Integration" OR "Knowledge Discovery" OR "Knowledge Organizational" OR "Knowledge Transference").

H. Inclusion criteria (ICi)

To address the research questions, the following criteria were defined:

[IC1] Processes, techniques, methods, practices and / or tools adopted to promote knowledge management in distributed software development;

[IC2] Scenarios and cases in which knowledge management has been applied to the distributed software development;

[IC3] Difficulties or problems found to implement knowledge management in distributed software development.

I. Exclusion criteria (ECi)

To address the research questions, the following criteria were defined:

[EC1] Processes, techniques, methods, practices and / or tools adopted to promote knowledge management those are not adequate for distributed software development context;

[EC2] Scenarios and cases in which knowledge management has not been applied for distributed software development;

[EC3] Difficulties or problems found in the implementation of knowledge management not related to distributed software development;

[EC4] Papers written in language different from English;

[EC5] Papers that were not available to perform the complete reading.

J. Preliminary Process

In the preliminary process of selection, the query string, composed by keywords and synonymous terms (Knowledge Management and Distributed Software Development), was submitted to indexed databases and search engines. Later, the title, abstract and keywords of found papers were read. Each paper selected was analyzed and was approved for complete reading if presented results that were relevant to research carried out. If there was not relevant data for research in question, the paper was not approved for complete reading.

K. Final Selection Process

At the end of selection process, the papers that had been approved in preliminary phase were undergone to a complete reading. So, when the subject (approach, tools, experiment) dealt was not related to research goal, the paper was discarded. At the end of the complete reading, if the paper was approved, the reader was responsible for summarize the data.

L. Quality assessment of the papers (Qi)

The assessment goal was not to classify the papers according to a total quality score. So, was not assigned any score. Thus to classify them the binary scale was used, considering the criteria defined by Dyba et al [5] as following:

[Q1] Is there a clear statement of the research goals?
[Q2] Is there an adequate description of the context in which the research was carried out?
[Q3] Does the research design is appropriate to address the aims established for that?
[Q4] Was there a control group with which could compare the results?
[Q5] Does the data analysis is sufficiently rigorous?
[Q6] Does the article is based on research or experiments?
[Q7] Are there results which can be applied in practical situations?
[Q8] Is there a clear statement of findings?
[Q9] Are there artifacts generated as result from the search presented at the paper?
[Q10] Does article presents some model for promoting the Knowledge Management?

3 Results and Discussion

After carry out the procedures defined in previous section, 38 primary studies were selected. Fig. 1 shows the method used and the results obtained at each stage.

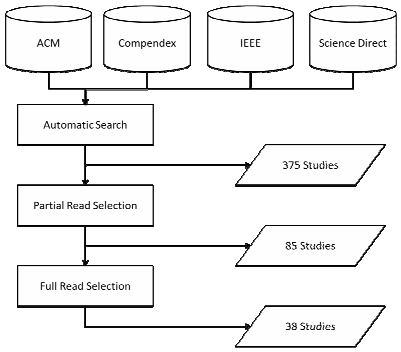


Fig. 1. Results of systematic review

In Table 1 are highlighted properties and characteristics concerned to knowledge management discussed in the selected papers (SP).

Table 1. Properties and Characteristics.

Properties and Features	Papers	Occurrence
Model	SP 1, SP 3, SP 4, SP 5, SP 8, SP 9, SP 10, SP 12, SP 13, SP 14, SP 15, SP 16, SP 17, SP 25, SP 26, SP 27, SP 29, SP 30, SP 33, SP 34	20 papers
Knowledge representation	SP 2, SP 3, SP 4, SP 5, SP 6, SP 7, SP 9, SP 11, SP 12, SP 13, SP 16, SP 17, SP 20, SP 21, SP 22, SP 23, SP 24, SP 25, SP 26, SP 27, SP 28, SP 30, SP 32, SP 35	24 papers
Knowledge dissemination	SP 10, SP 12, SP 17, SP 18, SP 19, SP 20, SP 23, SP 28, SP 34	9 papers
Knowledge flow	SP 1, SP 5, SP 7, SP 14, SP 28, SP 35	6 papers
Support to decision making	SP 3, SP 6, SP 7, SP 16, SP 21, SP 37	6 papers
Use of repository	SP 1, SP 2, SP 8, SP 14, SP 17, SP 19, SP 21, SP 24, SP 27, SP 28, SP 29, SP 30	12 papers
Use of contextual information	SP 4, SP 10, SP 13, SP 15, SP 25, SP 30, SP 31, SP 35	8 papers
Communication Standard	SP 6, SP 7, SP 8, SP 10, SP 11, SP 12, SP 28, SP 31, SP 37, SP 38	10 papers
Capturing knowledge/ experience	SP 2, SP 5, SP 7, SP 11, SP 17, SP 19, SP 20, SP 22, SP 25, SP 28, SP 29, SP 32, SP 34, SP 37	14 papers
Strategy for knowledge reuse	SP 14, SP 25	2 papers

Among the papers that presented models, 70% of these present a software for automation of them. The automation of the proposed model demonstrates its applicability and also facilitates its adoption.

Regarding the representation of knowledge was observed that most of them use resources such as templates and tags. Only one of the selected papers used ontology to promote knowledge management in distributed development. The ontology was used as a tool for standardization and quantification of the information and helped on externalization of knowledge. It is noteworthy that the benefits of using ontologies are related with the facility to carry out inferences and allow the reuse of information/knowledge. OntoDiSEN, is an ontology for distributed software development domain, which can be extended to include matters related to knowledge management [6].

The concept of reuse was also used as a justification of knowledge management, aiming at the classification of information and subsequent use of them.

E-mail, wiki, video conferencing and telephone are resources used for knowledge dissemination. However, it is emphasized that the use of appropriate tools to support knowledge dissemination is essential to ensure the knowledge availability on site and on time for correct decision making. In case of distributed software development the problems related to communication are exacerbated by socio-cultural and temporal questions.

The flow of knowledge was considered in 6 papers and demonstrated the importance of communication for development and maintenance of knowledge as intellectual capital owned by the company.

Regarding the use of repositories, the papers analyzed highlighted how important is to store the information concerned to the project, process, distributed teams and artifacts generated during whole development of distributed software. The papers also showed the importance of using of appropriated techniques to capture and represent of knowledge so that it can be processed and thus generate new knowledge.

In respect to communication standards, the papers presented a set of best practices, such as: knowing the sender to establish trust, standardization of documents to be sent, among others. Thus, adopting these best practices together with the formalization and use of automated tools could be a strategy to support knowledge reuse.

Studies point that culture has a great impact on promoting of Knowledge Management in DSD, since the information provided informally has higher interactivity than those obtained from tools and repositories. Based on this, it is emphasized that when dealing with distributed team, knowledge management cannot be disassociated from practices that promote the trust among team members. Another analysis was also performed to evaluate the year of paper publication. Fig. 2 shows 2010 as the year in which were found the most relevant publications presented in this review and also that there was no relevant publication in 2000 and 2001.

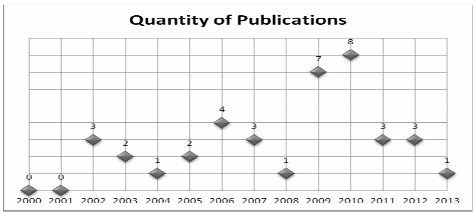


Fig. 2. Quantity of publications per year.

As for assessing the quality of work, the questions set out in methodology of the systematic review were answered. The results on Table 2, show the quantity of papers that met the particular issue of quality in the occurrence column.

Table 2. Quality Evaluation of Selected Papers.

	Question	Papers	Occurrence
Q1	Is there a clear statement of the research goals?	SP 1, SP 2, SP 3, SP 4, SP 5, SP 6, SP 7, SP 11, SP 12, SP 13, SP 15, SP 16, SP 17, SP 18, SP 19, SP 20, SP 21, SP 23, SP 24, SP 26, SP 28, SP 29, SP 30, SP 31, SP 32, SP 33, SP 34, SP 35, SP 36, SP 37, SP 38	33 papers
Q2	Is there an adequate description of the context	SP 2, SP 3, SP 4, SP 5, SP 7, SP 8, SP 9, SP 10, SP 11, SP 12, SP 15, SP 16,	27 papers

	in which the research was carried out?	SP 17, SP 18, SP 19, SP 20, SP 22, SP 23, SP 24, SP 25, SP 26, SP 27, SP 28, SP 31, SP 34, SP 35, SP 36, SP 37, SP 38	
Q3	Does the research design is appropriate to address the aims established for that?	SP 1, SP 2, SP 3, SP 4, SP 5, SP 7, SP 8, SP 10, SP 11, SP 12, SP 13, SP 15, SP 16, SP 17, SP 18, SP 19, SP 20, SP 23, SP 24, SP 26, SP 28, SP 29, SP 30 SP 31, SP 34, SP 35, SP 36, SP 37, SP 38	29 papers
Q4	Was there a control group with which could compare the results?	SP 7, SP 11, SP 12, SP 15, SP 17, SP 34	6 papers
Q5	Does the data analysis is sufficiently rigorous?	SP 2, SP 3, SP 5, SP 7, SP 11, SP 12, SP 15, SP 17, SP 19, SP 23, SP 24, SP 25, SP 26, SP 29, SP 30, SP 32, SP 35, SP 36, SP 38	19 papers
Q6	Does the article is based on research or experiments?	SP 1, SP 2, SP 3, SP 4, SP 5, SP 6, SP 7, SP 8, SP 9, SP 10, SP 11, SP 12, SP 13, SP 14, SP 15, SP 16, SP 17, SP 18, SP 19, SP 20, SP 21, SP 23, SP 25, SP 26, SP 27, SP 28, SP 29, SP 30, SP 32, SP 33, SP 34, SP 35, SP 36, SP 37, SP 38	36 papers
Q7	Are there results which can be applied in practical situations?	SP 1, SP 2, SP 3, SP 4, SP 7, SP 8, SP 11, SP 16, SP 18, SP 20, SP 24, SP 26, SP 29, SP 31, SP 32, SP 33, SP 34, SP 35, SP 37	19 papers
Q8	Is there a clear statement of findings?	SP 1, SP 2, SP 3, SP 4, SP 5, SP 6, SP 7, SP 8, SP 10, SP 11, SP 12, SP 13, SP 15, SP 16, SP 17, SP 18, SP 19, SP 20, SP 21, SP 23, SP 24, SP 25, SP 26, SP 28, SP 29, SP 30, SP 31, SP 32, SP 33, SP 35, SP 36, SP 37	32 papers
Q9	Are there artifacts generated as result from the search presented at the paper?	SP 1, SP 2, SP 3, SP 4, SP 9, SP 13, SP 16, SP 20, SP 24, SP 26, SP 27, SP 29, SP 31, SP 35	14 papers
Q10	Does article presents some model for promoting the Knowledge Management?	SP 1, SP 2, SP 3, SP 4, SP 5, SP 6, SP 8, SP 11, SP 24, SP 26, SP 27, SP 29, SP 31, SP 34, SP 35, SP 37, SP 38	17 papers

It can be observed that 50% of papers present results of knowledge management practices in distributed software development environment. Only 6 papers stand out practical experiments containing a control group for comparison and validation of hypotheses. Among these papers that present practices, most quote communication as a difficulty, thus emphasizing the need to establish the sense of trust in the sender of the message for the effective acquisition and internalization of knowledge reported.

4 Final Considerations

Regarding the studies analyzed, we observed that they address somehow, techniques, practices, tools and methods to promote knowledge management in distributed software development. Most studies showed knowledge representations and definitions for creating models of knowledge management, since there is specific knowledge in a DSD. Other studies showed the use of data repository to capture and maintain knowledge in the organization. To meet the challenge of communication in geographically dispersed teams, the use of e-mail, phone and web conferencing practices were more suitable.

This paper aimed to present the results achieved through a systematic review, listing the works that dealt with processes, techniques, methods, and practices to promote KM tools in DSD. Based on the analysis contained in Section III, we can highlight some points that deserve further study:

- appropriate mechanisms for disseminating knowledge;
- explore data mining techniques, observing socio cultural aspects in DSD, from data stored in the repository in order to extract interesting knowledge;
- mechanisms to capture and represent data, aiming to generate knowledge. One idea that has been explored by our research group is to proceed with the extension of OntoDiSEN [6];
- integrate knowledge management practices to approaches for distributed software development, such as proposed in [7]. It is important emphasize how artifacts can be used to disseminate knowledge in distributed teams [8];
- explore the context information as way for dissemination of knowledge.

Therefore, the results from this systematic review, evidence that there are challenges in knowledge management for DSD. So, the integration of several areas for standardization and reuse in software, storage and dissemination of information and communication are points that could be explored. Furthermore, the results show the need for tools that promote knowledge management in DSD to enable the socialization of information. For each one of these points we can foresee an interesting point for future works to be developed as research.

References

1. Herbsleb, J. D., Mockus, A., Finholt, T. A., Grinter, R. E.: Distance, dependencies, and delay in a global collaboration. In: ACM conference on Computer supported cooperative work (CSCW '00)', *Proceedings*…, USA, pp. 319--328, (2000)
2. Jimenez, M., Piattini, M., Vizcano, A.: Challenges and Improvements in Distributed Software Development: A Systematic Review. Advances in Software Engineering, vol. 2009, Article ID 710971, 14 pp. doi:10.1155/2009/710971 (2009)
3. Kitchenham, B.: Procedures for Performing Systematic Reviews. Joint Technical Report, Software Engineering Group, Department of Computer Science, Keele University, Empirical Software Engineering, National ICT Australia Ltd., Australia (2004)
4. Dyba, T., Dingsoyr, T., Hanssen, G.K.: Applying systematic reviews to diverse study types: An experience report. In: First International Symposium on Empirical Software Engineering and Measurement (ESEM), *Proceedings*… pp. 225--234. (2007)
5. Chaves, A. P., Steinmacher, I. F., Huzita, E. H. M., Leal, G. C. L., Biasao, A.B.: OntoDiSENv1: an Ontology to Support Global Software Development. CLEI Electronic Journal, vol. 14, pp. 1--12, (2011)

6. Leal, G. C. L., Chaves, A. P., Huzita, E. H. M., Delamaro, M. E.: An Integrated Approach of Software Development and Test Processes to Distributed Teams. Journal of Universal Computer Science (print), vol. 18, n. 19, pp. 2686--2705, (2012)
7. Kokkoniemi, J.: A Preliminary Model for Generating Experience Knowledge Based Artifacts. In: 39th Annual Hawaii International Conference on System Sciences (HICSS), *Proceedings*… DOI: http://doi.ieeecomputersociety.org/10.1109/HICSS.2006.25, (2006)

SELECTED PAPERS

[SP 1] Zhuge, H.: Knowledge flow management for distributed team software development. Knowledge-Based Systems, Elsevier, vol. 15, pp. 465--471, (2002)
[SP 2] Brandt, S. C., Morbach, J., Miatidis, M., Theien, M., Jarke, M., Marquardt, W.: An ontology-based approach to knowledge management in design processes. Computers and Chemical Engineering, vol. 32, n.1-2, pp. 320--342, (2008)
[SP 3] Karacapilidis, N., Adamides, E., Evangelou, C.: A computerized knowledge management system for the manufacturing strategy process. Computers in Industry, vol. 57, n. 2, pp. 178--188, (2006)
[SP 4] Ahn, H. J., Lee, H. J., Cho, K., Park, S. J.: Utilizing knowledge context in virtual collaborative work. Decision Support Systems, vol.39, n. 4, pp. 563--582, (2005)
[SP 5] Ho, C.-T., Chen, Y.-M., Chen, Y.-J., Wang, C.-B.: Developing a distributed knowledge model for knowledge management in collaborative development and implementation of an enterprise system. Robotics and Computer-Integrated Manufacturing, vol. 20, n. 5, pp. 439--456, (2004)
[SP 6] Fiore, S. M., Cuevas, H. M., Scielzo, S., Salas, E.: Training individuals for distributed teams: Problem solving assessment for distributed mission research. Computers in Human Behavior, vol. 18, n. 6, pp. 729--744, (2002)
[SP 7] Gupta, A., Mattarelli, E., Seshasai, S., Broschak, J.: Use of collaborative technologies and knowledge sharing in co-located and distributed teams: Towards the 24-h knowledge factory. Journal of Strategic Information Systems, vol.18, n. 3, pp. 147--161, (2009)
[SP 8] Pike, W., Gahegan, M.: Beyond ontologies: Toward situated representations of scientific knowledge. International Journal of Human Computer Studies, vol. 65, n. 7, pp. 674--688, (2007)
[SP 9] Alavi, M., Tiwana, A.: Knowledge integration in virtual teams: The potential role of KMS. Journal of the American Society for Information Science and Technology, vol. 53, n. 12, pp. 1029--1037, (2002)
[SP 10] Damian, D., Zowghi, D.: An insight into the interplay between culture, conflict and distance in globally distributed requirements negotiations. In: 36th Annual Hawaii International Conference on System Sciences, *Proceedings*..., DOI: http://dx.doi.org/10.1109/HICSS.2003.1173665, 10 pp., (2003)
[SP 11] Sarker, S., Sarker, S., Nicholson, D., Joshi, K.: Knowledge transfer in virtual information systems development teams: an empirical examination of key enablers. In: 36th Annual Hawaii International Conference on System Sciences, *Proceedings*…, DOI: http://dx.doi.org/10.1109/HICSS.2003.1174272, 10 pp., (2003)
[SP 12] Wu, S., Lin, C. S., Lin, T.-C.: Exploring knowledge sharing in virtual teams: A social exchange theory perspective. In: The Annual Hawaii International Conference on System Sciences, *Proceedings*…, vol. 1, pp. 26b, (2006)
[SP 13] Ye, Y.: Dimensions and forms of knowledge collaboration in software development. In: 12th Asia-Pacific Software Engineering Conference, *Proceedings*…, pp. 805--812, (2005)
[SP 14] Desouza, K. C., Awazu, Y., Baloh, P.: Managing knowledge in global software development efforts: Issues and practices. IEEE Software, vol. 23, n. 5, pp. 30--37, (2006)
[SP 15] Xu, B.: Enabling involving global cooperative software design with layered knowledge management platform. In: 11th International Conference on Computer Supported Cooperative Work in Design, *Proceedings*…, pp. 687--692, (2007)
[SP 16] Avram, G.: Of deadlocks and peopleware - Collaborative work practices in global software development. In: Second IEEE Intern. Conf. on Global Software Engineering, *Proceedings*…, pp. 91--102, (2007)
[SP 17] Boden, A., Avram, G.: Bridging knowledge distribution - the role of knowledge brokers in distributed software development teams. In: ICSE Workshop on Cooperative and Human Aspects on Software Engineering, CHASE 2009, *Proceedings*..., pp. 8--11, (2009)
[SP 18] Boden, A., Avram, G., M Bannon, L., Wulf, V.: Knowledge management in distributed software development teams - does culture matter? In: Fourth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 18--27, (2009)

[SP 19] Clerc, V., Lago, P., Van Vliet, H.: The usefulness of architectural knowledge management practices in GSD. In: Fourth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 73--82, (2009)
[SP 20] Taweel, A., Delaney, B., Zhao, L.: Knowledge management in distributed scientific software development. In: Fourth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 299--300, (2009)
[SP 21] Taweel, A., Delaney, B.; Arvanitis, T., Zhao, L.: Communication, Knowledge and Co-ordination Management in Globally Distributed Software Development: Informed by a scientific Software Engineering Case Study., In: Fourth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp.370--375, (2009)
[SP 22] Lee, S. B., Shiva, S. G.: A novel approach to knowledge sharing in software systems engineering. In: Fourth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…,pp. 376—381, (2009)
[SP 23] Beecham, S., Noll, J., Richardson, I., Ali, N.: Crafting a global teaming model for architectural knowledge. In: Fifth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 55--63, (2010)
[SP 24] Solis, C., Ali, N.: Distributed Requirements Elicitation Using a Spatial Hypertext Wiki. In: Fifth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 237--246, (2010)
[SP 25] Betz, S., Oberweis, A., Stephan, R.: Knowledge transfer in IT offshore outsourcing projects: An analysis of the current state and best practices In: Fifth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 330--335, (2010)
[SP 26] Salger, F., Sauery, S., Engelsy, G., Baumannz, A.: Knowledge transfer in global software development - Leveraging ontologies, tools and assessments. In: Fifth IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp.336--341, (2010)
[SP 27] De Boer, R. C., Van Vliet,: H. Experiences with semantic wikis for architectural knowledge management. In: 9th Working IEEE/IFIP Conf. on Software Architecture, WICSA, *Proceedings*…, pp. 32--41, (2011)
[SP 28] Clerc, V., Lago, P., Van Vliet, H.: Architectural knowledge management practices in agile global software development. In: 6th IEEE International Conference on Global Software Engineering ICGSE, *Proceedings*…, pp. 1--8, (2011)
[SP 29] Averbakh, A., Knauss, E., Liskin, O.: An experience base with rights management for global software engineering. In: 11th International Conference on Knowledge Management and Knowledge Technologies, *Proceedings*…, n. 10, (2011)
[SP 30] Clerc, V., De Vries, E., Lago, P.: Using wikis to support architectural knowledge management in global software development. In: International Conference on Software Engineering, *Proceedings*…, pp. 37--43, (2010)
[SP 31] Moraes, A., Silva, E., Da Trindade, C., Barbosa, Y., Meira, S.: Recommending experts using communication history. In: International Conference on Software Engineering, *Proceedings*…, pp. 41--45, (2010)
[SP 32] Noll, J., Beecham, S., Richardson, I.: Global software development and collaboration: Barriers and solutions. ACM Inroads, vol. 1, n. 3, pp. 66--78, (2010)
[SP 33] Salger, F., Engels, G.: Knowledge transfer in GSD: leveraging acceptance test case specifications. In: ACM/IEEE 32nd Intern. Conference on Software Engineering, *Proceedings*…, vol. 2, pp. 211--214, (2010)
[SP 34] Sengupta, B., Chandra, S., Sinha, V.: A research agenda for distributed software development. In: 28th International Conference on Software Engineering, *Proceedings*…, pp. 731--740, (2006)
[SP 35] Boden, A., Avram, G., Bannon, L., Wulf, V.: Knowledge sharing practices and the impact of cultural factors: reflections on two case studies of offshoring in SME. Journal of Software: Evolution and Process, vol. 24, n. 2, pp. 139--152, (2012)
[SP 36] Giuffrida, R., Dittrich, Y.: Empirical Studies on the Use of Social Software in Global Software Development - a Systematic Mapping Study. Information and Software Technology, vol. 55, n. 7, pp. 1143--1164, (2013)
[SP 37] Klitmoller, A., Lauring, J.: When global virtual teams share knowledge: Media richness, cultural difference and language commonality. Journal of World Business, vol. 48, n. 3, pp. 398--406, (2012)
[SP 38] Portillo-Rodriguez, J., Vizcano, A., Piattini, M., Beecham, S.: Tools used in Global Software Engineering: A systematic mapping review. Information and Software Technology, vol. 54, n. 7, pp. 663--685, (2012)

Propuesta de una Metodología para el Análisis de Adopción de Cloud Computing en PyMEs

Bernal, L.[1], Vegega, C.[1], Pytel, P.[1,2], Pollo-Cattaneo, M. F.[1]

[1] Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional. Argentina.

[2] Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina.

bernal.luciano@gmail.com; ppytel@gmail.com; fpollo@posgrado.frba.utn.edu.ar

Resumen. El concepto de Cloud Computing hace referencia a un modelo que permite habilitar acceso a la red, de forma conveniente y en demanda, a un fondo compartido de recursos computacionales configurables. Se ha observado la falta de una metodología homogénea que permita analizar la conveniencia y la viabilidad de la adopción de esta tecnología dentro de las Pequeñas y Medianas Empresas (PyMEs). Por lo tanto, el presente trabajo tiene como objetivo proponer una metodología que permita definir un proceso para analizar la conveniencia y la viabilidad de la adopción de la tecnología Cloud Computing dentro de las PyMEs. Para ello se tienen en cuenta no sólo los aspectos técnicos o económicos sino que se realiza un análisis integral de la estructura organizacional.

Palabras Claves: Cloud Computing. Metodología. Viabilidad y conveniencia. Análisis de adopción. PyMEs.

1. Introducción

Según el Instituto Nacional de Estándares y Tecnología de Estados Unidos (también conocido como NIST por su sigla en inglés), el concepto de Cloud Computing hace referencia a un "modelo que permite habilitar acceso a la red, de forma conveniente y en demanda, a un fondo compartido de recursos computacionales configurables (redes, servidores, almacenamiento, aplicaciones y servicios) que puede ser provisto rápidamente y con un mínimo esfuerzo de administración o interacción con el proveedor" [1]. También se puede agregar en este concepto, al hardware y los sistemas de software en los centros de datos que proveen los servicios entregados por demanda [2]. Estos servicios se denominan normalmente 'Software como Servicio' (o SaaS, por sus siglas en inglés), mientras que los recursos IT (hardware y software del centro de datos) necesarios, son lo que se llama 'Cloud' o 'la Nube'. Esta Nube se basa en la virtualización de recursos de

hardware, cuya comercialización se encuentra acompañada de sistemas de software que permiten gestionar la arquitectura subyacente. Por lo tanto, el paradigma de Cloud Computing ayuda a optimizar los procesos de almacenamiento y manejo de datos, haciendo más eficaz la toma de decisiones en una organización.
En [3] se ha observado la falta de una metodología homogénea que permita analizar la conveniencia y la viabilidad de la adopción de esta tecnología dentro de las Pequeñas y Medianas Empresas (PyMEs). Por lo tanto, el presente trabajo tiene como objetivo proponer una metodología que permita analizar la viabilidad de la adopción de Cloud Computing en PyMEs. Para ello, se define el concepto de PyME, con su respectiva clasificación (sección 2) y se propone la metodología junto con los distintos factores, tanto cualitativos como cuantitativos, que se deben evaluar para adoptar la tecnología de la Nube (sección 3). En la sección 4 se desarrollan los pasos de dicha metodología, con una prueba de concepto exitosa. Por último, en la sección 5 se detallan las conclusiones y futuras líneas de trabajo.

2. Características de las PyMEs

Las PyMEs constituyen el mayor sector empresarial de la actividad económica de América Latina y el Caribe, siendo en muchos casos el sector de movilidad de capital más importante para las economías nacionales [4].
En un sentido amplio, la PyME "es una unidad económica, dirigida por su propietario de forma personalizada y autónoma, de pequeña dimensión en cuanto a número de trabajadores y cobertura de mercado" [5]. Sin embargo, la definición de PyME no se encuentra estandarizada internacionalmente [6], algunos países las clasifican por volumen de ventas y otros, por cantidad de empleados [7]. El Fondo Multilateral de Inversiones (FOMIN), define a una PyME como aquella que tiene menos de 100 empleados y factura anualmente hasta US$ 3.000.000. En la definición general del MERCOSUR, el tamaño de la empresa, también queda definido bajo los dos criterios conjuntos: cantidad de empleados y ventas anuales. Sin embargo, se explicita que prevalece el de ventas, y el de cantidad de empleados es utilizado sólo como referencia. Aquí los límites de clasificación difieren de acuerdo al sector de actividad de pertenencia de la empresa, distinguiendo por un lado a la Industria y por otro lado a las de Comercio y Servicios [8]. Se observa que en los estados parte del MERCOSUR son utilizadas diversas definiciones para delimitar este universo que denota la heterogeneidad de criterios respondiendo a la naturaleza misma del fenómeno MPyMEs (Micro, Pequeñas y Medianas Empresas), que se origina y desenvuelve en distintas estructuras productivas. A su vez, las diferentes formas de acotar ese universo están en función de los objetivos que se persigue, a la precariedad de información y el contexto económico. Se las clasifica de la siguiente manera: aquellas pertenecientes al rubro de comercio y servicios, son empresas pequeñas de hasta 30 empleados, y medianas, de 31 a 80. Las ventas anuales de las pequeñas, no superan los U$S 1.500.000. En cambio, en las medianas, llegan hasta los U$S 7.000.000.

3. Metodología de Análisis de Adopción Propuesta

A diferencia de otros estudios sobre el tema, que sólo analizan los aspectos técnicos (ancho de banda, gastos de energía, almacenamiento entre otros) a la hora de adoptar el paradigma de Computación en la Nube [9] focalizándose en la viabilidad económica y rentabilidad [10], el presente trabajo plantea un análisis integral de la estructura organizacional de la PyME. Se tienen en cuenta tanto factores cualitativos (como la respuesta al cambio de los miembros de la organización, el grado en que los empleados trabajan en ubicaciones remotas, requisitos de seguridad y el nivel de estandarización de los procesos de la empresa), así como también los aspectos cuantitativos (que abarcan, desde la estimación temprana de beneficios en la Nube contra los beneficios en un centro de datos interno, el estudio de la infraestructura actual y el análisis de compatibilidad y portabilidad de aplicaciones).
En las subsecciones 3.1 y 3.2, se detallan ambos grupos de variables, para luego plantear las fases de la metodología en la subsección 3.3.

3.1 Aspectos Cualitativos

Los aspectos cualitativos a contemplar para migrar a entornos Cloud Computing son:

- *Tipo de Aplicaciones*
  Es prioritario identificar, antes de abordar un modelo Cloud, qué es lo que se puede migrar a este tipo de servicio y qué no. IBM [11] discrimina dos tipos de carga de trabajo según la arquitectura de servicio:
  Para ingresar a un Cloud Público las cargas de trabajo de infraestructura suelen ser las más apropiadas:
  - Servicio de Mesa de Ayuda (Help Desk).
  - Infraestructura para entrenamiento y demostración.
  - Infraestructura de Video y Voz sobre IP (también conocida como Infraestructura WoIP).
  - Servidores.

  Mientras que para ingresar a un Cloud Privado, las bases de datos y aplicaciones de trabajo suelen ser las más apropiadas:
  - Minería de Datos y Minería de Texto.
  - Data-warehouses y data-marts.
  - Archivos con información de largo plazo.
  - Bases de datos transaccionales.

- *Requisitos de Cumplimiento normativo*
  Según [12] se pueden aplicar requisitos normativos estrictos para algunos datos, como puede ser la información financiera o, en salud, ciertos datos de los pacientes

como las historias clínicas. Las leyes y normas obligan a la estandarización de los procedimientos organizacionales. Las PyMEs tienden a aumentar el nivel de normalización de sus procesos [13]. Por tal motivo, es vital tener en cuenta cuál es la información estandarizada, ya que ésta será la más adecuada para migrar a la Nube.
Se debe tener en cuenta que la organización más apta para adoptar el paradigma Cloud Computing, es aquella abierta, basada en estándares y orientada a servicios.

- *Documentación y Medición de la Infraestructura Actual*
  Antes de considerar un modelo Cloud es fundamental tener en cuenta cuáles son las características actuales y potenciales de la infraestructura IT de la empresa. Se deberá contar con una infraestructura dinámica, con capacidades de virtualización, y provisión automatizada de recursos. Una infraestructura preparada para un modelo Cloud se caracteriza por poseer una virtualización avanzada y gestión automatizada, cuya infraestructura tiene un nivel de seguridad avanzada.

- *Acuerdo de Nivel de Servicios*
  Como la adopción de una arquitectura Cloud requiere la contratación de una entidad que brinde esa arquitectura y los servicios que la acompañan, es posible considerarla básicamente como la tercerización de una parte de la organización. Las empresas de mayor tamaño suelen poseer más experiencia en la documentación y protocolos requeridos de este tipo de proceso. Como las PyMEs, proporcionalmente, suelen derivar menos funciones a terceros, es necesario realizar un análisis riguroso del contenido del Acuerdo de Nivel de Servicios, o SLA por sus siglas en inglés [14].
  El SLA es el documento que describe la relación entre el proveedor del servicio y el usuario. Es esencial a la hora de interactuar con un proveedor por contener una definición completa y precisa sobre cada servicio brindado y las responsabilidades de cada parte.

- *Factores técnico/operativos: modelos de arquitectura*
  En [2] se menciona que los modelos propuestos de arquitectura varían de acuerdo al proveedor, generando ambientes heterogéneos que hacen compleja la interconexión Inter Cloud. Debido a la falta de estandarización de las plataformas de desarrollo y las arquitecturas de administración de los recursos Cloud, se deben re-analizar los modelos de comercialización y licencias del software como servicio.
  En primer lugar, se debe hacer una discriminación entre los distintos participantes que representarían la "cadena de desarrollo" del Cloud Computing, partiendo del proveedor de la arquitectura, pasando por el proveedor del software como servicio y, por último, el usuario final. En la actualidad existen grandes compañías que aparecen

a lo largo de toda la secuencia. Son poseedoras de centros de datos, que luego de virtualizados, son vendidos, a través de distintos modelos de comercialización, como el "pay-as-you-go" (pago por uso/en demanda). A su vez, estas empresas ofrecen la administración de sus recursos de infraestructura (IaaS), para garantizar escalabilidad y flexibilidad de los mismos, y diseñan plataformas de desarrollo de aplicaciones (PaaS) con sus propios protocolos, que difieren de los de otras compañías, creando este "estado de naturaleza" y falta de estandarización, mencionados previamente.

Una ventaja para el proveedor de Cloud Computing podría ser acaparar toda la información de un cliente, sin que éste tenga la posibilidad de transferir los datos de un proveedor a otro, o que esto le represente al usuario mucho esfuerzo y dinero. Sin embargo, este último corre riesgo de entrar en un modelo monopólico de precios (se encuentra atado al aumento de precios que el proveedor disponga, cuando este lo determine).

- *Factores técnico/operativos: administración de los recursos e infraestructura*

  La oferta de otros servicios que crean valor agregado y aumentan la experiencia del usuario se relaciona más con el tipo y la calidad del servicio que con el factor costo. Es decir, muchas veces interesa más analizar el tipo de servicio que ofrece el proveedor, antes que el costo de dicho servicio.

  Considerando las distintas clases de administración de recursos (Utility Computing) dependiendo de las arquitecturas propuestas [15], en [2] se establece un espectro de arquitecturas posibles que tiene en un extremo a EC2 de Amazon mientras que en la otra están Google AppEngine y Force.com, pasando por varias plataformas intermedias como Microsoft Azure.

  Por un lado existen arquitecturas que se consideran más cerca de la máquina, por requerir "solo unas pocas docenas de llamados a la API para configurar el hardware virtualizado". Así se obtiene mayor facilidad y libertad con la que el programador puede desenvolverse en la arquitectura de AWS (Amazon Web Services).

  Por otro lado, AppEngine de Google y Force.com de SalesForce tienen mecanismos de auto escalabilidad, y disponen de una administración rigurosa sobre el almacenamiento y en cuánto uso de CPU que puede dedicarse a un pedido particular.

  Por último, existen servicios que soportan la ejecución de varios tipos de aplicaciones, permitiendo al programador configurar ciertas propiedades de la plataforma pero dejando otras fuera de su alcance. Es el caso de Microsoft Azure, que se encuentra en un punto intermedio del espectro.

  Por lo tanto, queda bastante claro que las arquitecturas de cada proveedor son las que determinan la clase de administración de recursos que ofrecen. Y esto se debe ajustar a la necesidad de cada cliente. Empresas que requieren rápida escalabilidad y

disposición de recursos automática (junto con un entorno de desarrollo prediseñado) pueden optar por el marco de aplicaciones de AppEngine por ejemplo, mientras que otras que necesitan administrar el hardware con mayor libertad, podrían elegir las máquinas virtuales de EC2. Por consiguiente es posible decir que aunque el precio de uno u otro servicio varíen notoriamente, es la necesidad específica de cada cliente, la que termina definiendo la elección de un proveedor.

3.2 Aspectos Cuantitativos - Estimaciones Económicas

En todo estudio de factibilidad es necesario llevar a cabo un análisis económico del proyecto. Este tipo de estimaciones toman como índice principal la Tasa de Retorno de la Inversión (TIR). Se basa, principalmente, en estimar el tiempo en que la inversión reporta beneficios teniendo en cuenta las utilidades que generará el resultado del proyecto y los costos de inversión del mismo.

Cuando se contrata un servicio de computación en la Nube, el oferente proporciona una herramienta de análisis de presupuestos. Por ejemplo, tanto la herramienta Windows Azure TCO de Microsoft [16] como la RDS Cost Comparison Calculator de Amazon [17] permiten estimar los costos operacionales de los servicios que ofrecen las empresas mencionadas y determinar los beneficios económicos generados debido a la diferencia entre el mantenimiento de un centro de datos privado y los precios de los servicios en la Nube. Estas herramientas se basan en el análisis de tres grandes factores que define el usuario: el ancho de banda a consumir, la cantidad de horas de uso de CPU´s, y el espacio de almacenamiento en discos.

Sin embargo, debido a que la cantidad de proveedores de computación en la Nube se incrementa día a día y que cada uno de ellos establece su propio modelo de costos, bajo su propia arquitectura, queda fuera del foco de este trabajo, el estudio del modelo de costo de cada proveedor existente en el mercado. Por lo tanto, contando sólo con los precios de la computación en la Nube para un proveedor dado, se puede recurrir a la siguiente fórmula genérica definida en [2], para lograr una estimación de costos temprana:

$$\text{HorasUsuario}_{\text{cloud}} \bullet (\text{Ingreso} - \text{CostoServicio}_{\text{cloud}}) \geq \text{HorasUsuario}_{\text{datacenter}} \cdot \bullet \left(\text{Ingreso} - \frac{\text{Costo}_{\text{datacenter}}}{\text{Utilización}} \right) \quad (1)$$

En la parte izquierda de la fórmula (1) se obtiene como resultado la ganancia esperada del uso de Cloud Computing (considerando el ingreso obtenido, el costo del servicio y las horas utilizadas) mientras que en la parte derecha se desarrolla el mismo cálculo para un centro de datos de una determinada capacidad (factorizándolo por el promedio de utilización para incluir los períodos en los que la carga de trabajo no presenta picos). Entonces si el valor del lado derecho (costo del centro de datos) es mayor valor al

izquierdo (costo de uso de la Nube), significa que existe la oportunidad de generar mayores beneficios adoptando dicha tecnología.

3.3 Fases de la Metodología Propuesta

La Metodología propuesta consta de las siguientes fases que se deben realizar en el orden recomendado:

A. Estudiar la integración de la arquitectura actual con el entorno Cloud Computing
Se debe detallar de forma precisa la arquitectura actual de la empresa. A través de la documentación existente se debe evaluar la compatibilidad y portabilidad de las plataformas y aplicaciones que se desean migrar. Esta fase constituye un primer filtro, para aquellas organizaciones que no reúnen las características necesarias de virtualización avanzada, gestión automatizada y un nivel de seguridad avanzada.

B. Planificar la administración de recursos humanos y procedimientos
Esta etapa es fundamental para mantener un flujo de comunicación constante dentro de la organización. En [18] se aclaran los beneficios de informar a toda la empresa de los alcances y consecuencias que el cambio producirá. Allí se aclara que la administración de los proyectos de software comienza con un conjunto de actividades de manera colectiva que constituyen la planificación. Se deben estimar el trabajo que se realizará, los recursos requeridos, y el tiempo que consumirá. Encarar un proyecto de este tipo no solo involucra una planificación detallada en la administración de los recursos humanos, sino que también afecta a los procesos que se llevan a cabo en la organización. Dado que la adopción de Cloud Computing tiende a mejorar la normalización de procedimientos orientados a la calidad, en el contexto específico de las PyMEs se logra orientarlas hacia procesos más estandarizados.

C. Elegir aplicaciones con riesgo y carga de trabajo bajos
Se debe discriminar entre las aplicaciones que se desean migrar a la plataforma en la Nube, a través de la clasificación propuesta en Tipo de Aplicaciones (sección 3.1). El contenido de la información que ellas administren debe ser analizado para luego evaluar el riesgo que conlleva virtualizar la arquitectura en donde se almacena.

D. Estimación Económica Temprana
Incluye el análisis de los factores cuantitativos propuestos anteriormente en la sección 3.2 y la implementación de la ecuación (1) para realizar una "estimación temprana". Aquí no se tienen en cuenta los precios de cada servicio que cada proveedor determina pero se indica la viabilidad económica del proyecto.

E. Desarrollo del SLA

Esta etapa se encuentra antes que la selección del proveedor ya que la PyME (cliente de los servicios cloud) debería dejar asentado en forma genérica sus requisitos técnicos, operativos y económicos. Este documento genérico, luego será adaptado para adecuarse a las particularidades del proveedor elegido.

F. Selección del proveedor

Teniendo en cuenta lo planteado en los factores técnico/operativos en la sección 3.1, se debe realizar el análisis de infraestructura y los modelos de arquitectura que cada proveedor ofrece. Con cada tipo de entrega de servicios se debe estudiar, el modelo de costos del proveedor.

4. Prueba de Concepto

Para validar la metodología para el análisis de adopción de Cloud Computing propuesta, se utiliza una prueba de concepto positiva con un proyecto real finalizado con éxito [19].
El objetivo que perseguía el Instituto Argentino de Responsabilidad Social Empresaria (IARSE) era afrontar en tiempo y forma la comunicación entre sus empresas miembros [20]. Entre las posibles soluciones se encontraba la adopción de Software como Servicio (modelo de entrega de servicios de Cloud Computing), o la compra de una aplicación y la consecuente compra de hardware para almacenar el tráfico de datos que la misma generaría.
Teniendo en cuenta la ubicación geográfica, la infraestructura de hardware de la organización (como lo indica la fase A de la metodología) y que los usuarios de la aplicación a implementar se encuentran a lo largo de toda la Argentina, se decidió analizar la adopción de Cloud Computing. Para ello, se llamó a los directores de área del IARSE y se informó del proyecto. Se llegó a la conclusión que, como la organización no contaba con una aplicación propia para compartir información entre los usuarios, se debía adquirir una externa. La misma trataría de una plataforma colaborativa de intercambio de información en tiempo real. Por lo que los datos que administraría no eran de alto riesgo (análisis correspondiente a la fase C).
A través del cálculo de costos y, antes de elegir un proveedor determinado, se evaluó la viabilidad económica del proyecto (fase D). Como resultado, se identificó que la migración a servidores virtualizados representaba, aproximadamente, un 200% de ahorro con respecto a la compra de hardware propio. Los directivos de IARSE decidieron elaborar una lista de servicios y funcionalidades que la aplicación y el oferente de servicios debía brindar. Lo que se asemeja a la creación del modelo genérico del nivel de servicios propuesto en la fase E.
Por último, luego del análisis de varias alternativas en base al costo, el valor agregado por servicios y los modelos de arquitectura ofrecidos por los distintos proveedores, se decidió la adopción de la tecnología LotusLive de IBM [21]. LotusLive es una herramienta de software de colaboración en modalidad cloud computing que le permite a IARSE utilizar

aplicaciones para Comunidad, Web Meetings y Foros de Discusión. Como la misma está alojada en bases de datos externas, no requiere gastos adicionales en lo que respecta a su mantenimiento. Esta solución resuelve la problemática de comunicación de la organización, basándose en el Software como Servicio (SaaS) y la virtualización de servidores. Actualmente, IARSE sigue utilizando la aplicación después de 2 años de su implementación y de la adopción de Cloud Computing.
Vale aclarar que, en el proyecto descripto anteriormente, la metodología no se implementa de forma rigurosa. Se observa que el paso B, no se lleva a cabo con la importancia que debería, ya que no solo se debe informar del cambio a los directores de área, como se describe, sino también al personal que interactuará con la nueva tecnología. Además, no se formalizó la documentación referida a la estandarización de procedimientos. A pesar de esto, este caso de éxito enfocó su estrategia de adopción de una manera muy similar a la que se propone en este trabajo.

5. Conclusión

El presente trabajo tiene como objetivo proponer una metodología que permita definir un proceso para analizar la conveniencia y la viabilidad de la adopción de la tecnología Cloud Computing dentro de las PyMEs. Para ello se proponen seis fases que consideran no sólo los aspectos técnicos o económicos sino que se realiza un análisis integral de la estructura organizacional. Como se puede observar, el estudio de viabilidad económica representa sólo uno de los factores que hay que analizar antes de implementar Cloud Computing. Es indispensable tener en cuenta las características propias de las PyMEs, la problemática y contexto propio de cada una para llevar a cabo un estudio más maduro si se piensa adoptar este nuevo paradigma.
Por último se ha analizado un caso de adopción de esta tecnología donde se puede observar que se ha llegado al éxito del proyecto por haber aplicado etapas similares a las fases propuestas en esta metodología.
Como futura línea se trabajará en la definición detallada de cada una de las fases propuestas indicando para cada una, las actividades que se deben realizar junto con las técnicas y herramientas recomendadas.

Referencias

1. Mell, P., Grance, T. (2011) The NIST Definition of Cloud Computing. Special Publication, National Institute of Standards and Technology
2. Armbrust, M., Fox, A. (2009) Above the Clouds: A Berkeley View of Cloud Computing. Technical Report, Electrical Engineering and Computer Sciences, University of California at Berkeley

3. Khajeh-Hosseini, A., Greenwood, D., Smith, J., Sommerville, I. (2010) The Cloud Adoption Toolkit: Supporting Cloud Adoption Decisions in the Enterprise.
4. Frankin, R.C., Pessoa de Matos, M. (2011) Apoyando a las PyMEs: Políticas de fomento en América Latina y el Caribe. CEPAL, Publicación de Naciones Unidas.
5. Cardozo, E., Velásquez de Naime, Y., Rodríguez Monroy,C. (2012) El concepto y la clasificación de PyME en América Latina. En: Global Conference on Business and Finance Proceedings.
6. Ueki, Y., Tsuji, M., Olmos, R.C. (2005) Tecnología de la información y las comunicaciones (TIC) para el fomento de las PyMEs exportadoras en América Latina y Asia oriental. CEPAL, Publicación de Naciones Unidas.
7. Fundación Observatorio PyME (2013) Informe Especial: Definiciones de PyME en Argentina y el resto del mundo. http://goo.gl/0VrU1
8. Políticas de apoyo a las micro, pequeñas y medianas empresas del MERCOSUR – Etapa II. En: XXXII GMC, MERCOSUR/GMC/RES. N° 59/98 (1998)
9. Espino Barrios, L. (2009) Cloud Computing como una red de servicios. Reporte Técnico, Instituto Tecnológico de Costa Rica.
10. Microsoft Corporation (S/A) The Economics of the Cloud. http://goo.gl/B1FpD
11. Fourcade, G. (S/A) Seis variables para analizar antes de saltar a la nube. IBM. http://goo.gl/nXotK
12. VMware, Inc. (2013) El momento adecuado para adoptar la virtualización de escritorios: Siete indicadores clave. http://goo.gl/W9Qxd
13. Orlandi, P. (S/A) Las PyMEs y su rol en el Comercio Internacional. White Paper Series del Centro de Estudios para el Desarrollo Exportador – CEDEX.
14. Cloud Computing Use Case Discussion Group (S/A) Cloud Computing Use Cases – Version 4.0. White Paper.
15. Vaquero, L.M., Rodero-Merino, L., Caceres, J., Lindner, M. (2009) A Break in the Clouds: Towards a Cloud Definition. Nota Editorial, ACM SIGCOMM Computer Communication Review, Vol. 39, N. 1.
16. Microsoft Corporation (S/A) Plataforma Windows Azure: TCO and ROI Calculator. http://goo.gl/26Apo
17. Amazon Web Services (S/A) RDS Cost Comparison Calculator http://goo.gl/pFiWU
18. Pressman, R.: (2004) Software Engineering A Practitioner´s Approach. Editorial Mc Graw Hill.
19. IBM (S/A) IARSE optimiza su comunicación y amplía el debate sobre RSE en Argentina a través de una plataforma colaborativa cloud de IBM. http://goo.gl/DXCCH
20. Rolando De Serra, A. (2012) IARSE - Construyendo un nuevo horizonte de RSE en Argentina (c.01 de lo periférico a lo estratégico). http://goo.gl/EOgrN
21. IBM (S/A) Software como servicio: LotusLive. http://goo.gl/5Bgop

Modelo para aplicaciones sensibles al contexto (MASCO): Un caso de estudio para validación.

Evelina Carola Velazquez[1], Ariel Nelson Guzman Palomino[1], María del Pilar Galvez Díaz[1], Nélida Raquel Caceres[1]

Universidad Nacional de Jujuy – Facultad de Ingeniería, San Salvador de Jujuy, Jujuy, Argentina
{mdpgalvezdiaz,nrcaceres}@fi.unju.edu.ar

Resumen. Este trabajo presenta la implementación de un caso de estudio para controlar en forma automatizada el funcionamiento de un invernadero utilizando el modelo MASCO para aplicaciones sensibles al contexto. El invernadero incluye sensores y actuadores y los valores de las variables de contexto consideradas para un normal desarrollo del cultivo presentan dependencia. Se realiza un trabajo de simulación, se obtiene una formula matemática para manejar la interdependencia de las variables de contexto y se utilizan patrones de diseño para abordar la complejidad. Como conclusión los patrones de diseño aplicados permiten mantener la integridad estructural y la característica de flexibilidad del modelo.

Palabras Clave: Ingeniería de Software - Modelos - Patrones - Context Aware

1 Introducción

Las aplicaciones sensibles al contexto permiten determinar lo que ocurre entre el sistema y su entorno, determinando qué, cómo y cuándo se presentan eventos externos a los que el sistema debe responder. En este contexto, se presenta la implementación de un proceso de control automático del clima de un invernadero, utilizando el Modelo MASCO, con el fin de determinar la adaptabilidad del modelo al caso propuesto y la flexibilidad del mismo para alcanzar un estado estable en el sistema. En el apartado 2 se describe el modelo MASCO, en el apartado 3 se describe el caso de estudio: Invernadero, en el apartado 4 se especifican las características de la implementación realizada y los patrones de diseño utilizados para dar solución a la toma de decisiones, en el apartado 5 se presentan las conclusiones y en el apartado 6 presentan las referencias.

2 Modelo MASCO

El Modelo MASCO (Modelo que provee servicios para aplicaciones sensibles al contexto) tuvo su origen en una extensión del modelo presentado en Gordillo [1] que considera servicios sensibles a la variable de contexto ubicación y al perfil del usuario

tomando como referencia el framework Context Toolkit basado en Widgets [2] y el modelo de automatización CIM -Computer Integrated Manufacturing- [3] que constituye un modelo de referencia que da soporte a aplicaciones industriales.

En las aplicaciones sensibles al contexto, donde existe más de una variable de contexto, el sistema debe gestionar diversos comportamientos u ofrecer diversos servicios en base al cambio de valor o de estado de una ó más variables de contexto o de su combinación. Además un objeto entidad o un objeto variable de contexto de la aplicación puede tener que relacionarse con uno ó más objetos que representan cada uno una variable de contexto o entidad. [4]

El hardware para el sensado evoluciona constantemente. Las reglas de sensado varían de acuerdo a las capacidades del hardware. Los datos sensados deben ser interpretados y la aplicación debería aparecer transparente a estos procesos. Esto se logra desacoplando los sensores de su lógica y lo concerniente a la aplicación [1]. MASCO, que contempla todas estas situaciones planteadas, se presenta en la Fig. 1, las áreas sombreadas representan los componentes presentados en [1] y las áreas punteadas las modificaciones realizadas a través de los trabajos presentados en [4], [5], [6], [7].

MASCO es un modelo en capas, donde se identifican cinco capas que se describen a continuación: [6]

- Aplication Layer: se encuentran los objetos del dominio de la aplicación.
- Context Layer: contiene los objetos necesarios para procesar la información de contexto.
- Service Layer: contiene los objetos necesarios para proveer servicios tanto internos como externos al sistema.
- Sensing Concern Layer: se encarga de interpretar o traducir los datos que provienen de Hardware Abstractions Layer.
- Hardware Abstractions Layer: en esta capa se agrupan los objetos que representan los sensores y actuadores.

3 Caso de estudio: Invernadero

El caso de estudio corresponde a la implementación de un sistema de control climático para un invernadero cuyos procesos se encuentran automatizados, el cual realiza el seguimiento de los parámetros de interés a través de sensores, comprueba las condiciones ambientales internas del invernadero en base a los valores sensados y las corrige utilizando actuadores sobre dispositivos automáticos instalados (ventanas laterales, riego por goteo, etc.), de manera que las condiciones climáticas sean óptimas para el correcto desarrollo y crecimiento de los cultivos allí ubicados. Para esto se establece un único proceso que consiste en el control de los parámetros de interés centralizados en el monitoreo de las cuatro variables principales del proceso fotosintético de una planta, las cuales constituyen variables de contexto para el modelo MASCO que deben comprobarse y corregirse si se presentan valores anómalos. Estas variables son: luminosidad, temperatura, dióxido de carbono y humedad relativa.

Luminosidad: Es la cantidad de radiación que es proyectada por una fuente de energía. En el caso de estudio es provista por el sol o por una fuente artificial que asegura que la planta reciba la cantidad de radiación necesaria para optimizar el proceso de fotosíntesis.

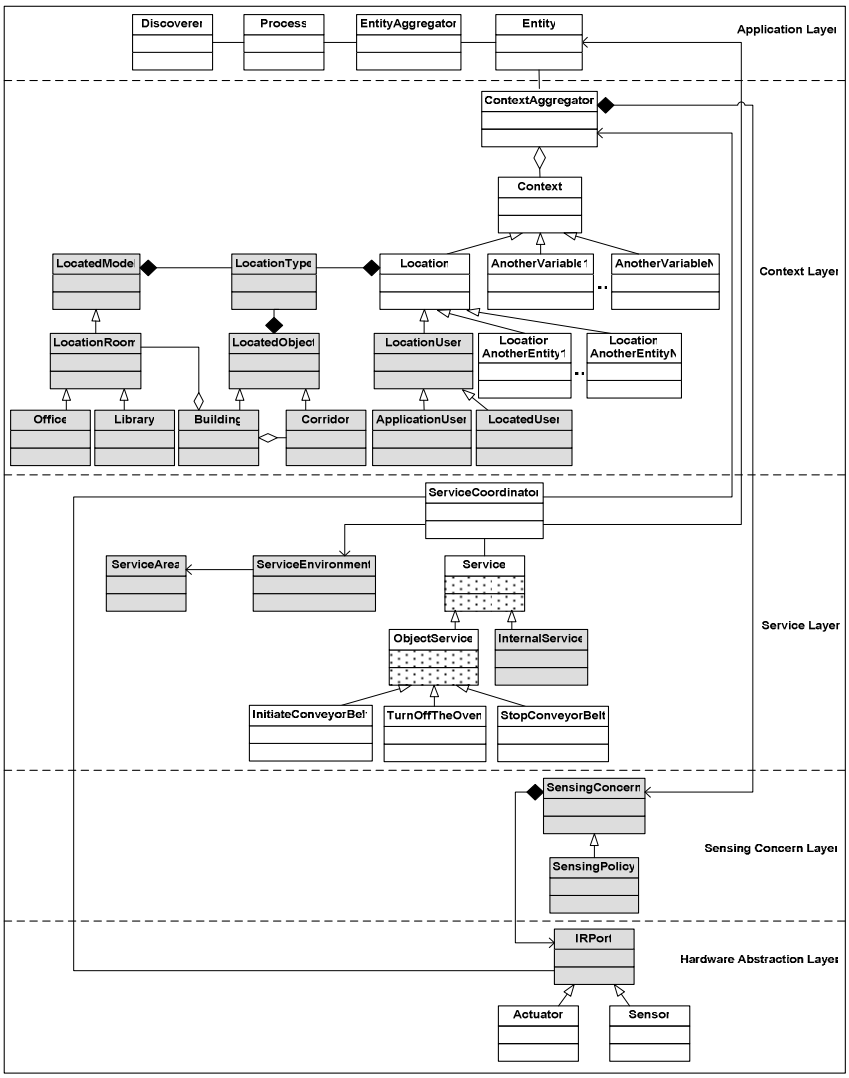


Fig. 1. Modelo MASCO

Temperatura: esta magnitud influye en el crecimiento y desarrollo de las plantas. La temperatura óptima general para las plantas se encuentra entre los 10 y 20º C. Para su manejo es importante conocer las necesidades y limitaciones de la especie cultivada. Se deben tener en cuenta además los valores que se deben alcanzar para el crecimiento óptimo del cultivo y sus limitaciones:

- Temperatura mínima letal: Aquella por debajo de la cual se producen daños en la planta.
- Temperaturas máximas y mínimas biológicas: Indican valores, por encima o por debajo respectivamente, de los cuales no es posible que la planta alcance una determinada fase vegetativa, como floración, fructificación, etc.
- Temperaturas nocturnas y diurnas: Indican los valores aconsejados para un correcto desarrollo de la planta.

La temperatura en el interior del invernadero se encuentra en función de la radiación solar, comprendida entre 200 y 4000 W/m^2; cuando la radiación solar es insuficiente para mantener la temperatura necesaria se toma la decisión de activar la fuente artificial de energía respetando los valores requeridos de las demás variables de contexto. Estableciendo una relación de dependencia entre la Temperatura y la Luminosidad.

Dióxido de Carbono: el anhídrido carbónico (CO_2) de la atmósfera es la materia prima imprescindible de la función clorofílica de las plantas. La concentración normal de CO_2 en la atmósfera es del 0,03%. Este índice debe aumentarse a límites de 0,1-0,2%, cuando los demás factores de la producción vegetal sean óptimos, si se desea el aprovechamiento al máximo de la actividad fotosintética de las plantas. Las concentraciones superiores al 0,3% resultan tóxicas para los cultivos. Los niveles aconsejados de CO_2 dependen de la especie o variedad cultivada, la radiación solar, la ventilación, la temperatura y la humedad. El óptimo de asimilación está entre los 18°C y 23º C de temperatura, descendiendo por encima de 23-24º C. Respecto a la luminosidad y humedad, cada especie vegetal tiene un valor óptimo distinto.

Sin embargo, no se puede hablar de una buena actividad fotosintética sin una óptima luminosidad. La luz es factor limitante, y así, la tasa de absorción de CO_2 es proporcional a la cantidad de luz recibida, además de depender también de la propia concentración de CO_2 disponible en la atmósfera de la planta. Se puede decir que el periodo más importante para el enriquecimiento carbónico es el mediodía, ya que es el momento en que se producen las máximas condiciones de luminosidad. Aquí también se establece una relación de dependencia entre la concentración de Dióxido de Carbono y la Luminosidad.

Humedad Relativa (HR): la humedad es la masa de agua en unidad de volumen o en unidad de masa de aire, cantidad de agua contenida en el aire, en relación con la máxima que sería capaz de contener a la misma temperatura. Una característica importante para el caso de estudio es la relación inversa entre la Temperatura y la Humedad Relativa: si la temperatura es elevada disminuye la HR, caso contrario la HR aumenta, lo que implica encontrar el equilibrio entre ambas cantidades para optimizar la fotosíntesis de la planta.

La Humedad Relativa del aire es un factor climático que puede modificar el rendimiento final de los cultivos. Cuando es excesiva las plantas reducen la transpiración y disminuyen su crecimiento, se producen abortos florales por apelmazamiento del polen y un mayor desarrollo de enfermedades criptogámicas. Por el contrario, si es muy baja, las plantas transpiran en exceso, pudiendo deshidratarse, además de los problemas en el cuajado.

3.1 Relación entre las variables de contexto

Al analizar el comportamiento de las variables de contexto consideradas para el caso de estudio, se observó una fuerte vinculación entre ellas, esto determina la condición del sistema de control que maneja la interacción de variables para mantener las condiciones estables del invernadero, esta vinculación es necesaria para lograr una fotosíntesis exitosa en las plantas. Esta interacción y la manera de realizar su control es motivo de estudio en este trabajo.

Relación vinculante de las variables de contexto:

- Luminosidad:
  - Temperatura. (Proporcionalmente).
  - Dióxido de Carbono. (Proporcionalmente).
  - Humedad Relativa. (Proporcionalmente).
- Temperatura:
  - Dióxido de Carbono. (Proporcionalmente).
  - Humedad Relativa. (Inversamente proporcionalmente).

Se puede generalizar esta vinculación de la siguiente forma:

$$t = f(l) \quad (1)$$

Donde t (temperatura) es función f de l (luminosidad) y:

$$c = g(t,l) \equiv c = g(f(l), l) \equiv c = g(l) \quad (2)$$

Donde c (dióxido de carbono) es función g de t (temperatura) y l (luminosidad), pero teniendo en cuenta que t también es función de l podemos concluir que dióxido de carbono depende de l y t, pero con mayor prioridad depende de l.

$$h = h(t,l) \equiv h = (f(l), l) \equiv h = h(l) \quad (3)$$

Donde h (humedad relativa) es función h de t (temperatura) y l (luminosidad), ya que t depende de l, entonces h depende de l y t, pero con mayor prioridad de l.

De (1), (2) y (3) deducimos que existe un orden prioritario de efectos entre las variables de contexto que poseen vinculación: Luminosidad, Temperatura, Dióxido de Carbono y Humedad Relativa.

3.2 Sensores

Para realizar el monitoreo de los valores de las variables de contexto se utilizan valores simulados de sensores en un período de tiempo predefinido, que depende del tiempo en el cual se realiza un cambio en el valor de la misma que sea representativo para el análisis y la apreciación del experto. La toma de valores se produce en el mismo período para las cuatro variables de contexto a fin de combinarlas para su evaluación y posterior decisión de acción correctiva. Se utilizaron sendos tipos de sensores para monitorear los valores de las cuatro variables de contexto, para el presente trabajo se considera la utilización de un sensor por cada una.

Se establece que la entidad a considerar es una planta de una especie en particular.

3.3 Actuadores

Los Actuadores son dispositivos de manejo automático de objetos, que corrigen los valores anómalos de una o más variable de contexto, así por ejemplo, una ventana lateral, que es un objeto para disminuir la temperatura dentro del invernadero, posee uno o más actuadores, motores, que permiten su apertura o cierre regulado. Para el caso de estudio, los actuadores corresponden a motores, servomotores y demás dispositivos que comandan los objetos que son utilizados para corregir los valores:

- Temperatura: Ventanas Laterales, Ventilador Refrigerante, Papel Refrigerante, Ventilador Calefactor y Bomba asociada a un termo tanque para levantar la temperatura.
- Humedad Relativa: Bomba electroválvulas, Pico de Riego, Riego por goteo.
- Iluminación: Fotocélula con reloj interno del controlador y grupo de luces.
- Concentración de Dióxido de Carbono: Compuertas de ventilación.

El control del ambiente del invernadero se realiza simulando el proceso de sensado, generando valores para las cuatro variables de contexto a partir de funciones que respetan su relación de dependencia, así por ejemplo, para la variable luminosidad, se generan valores de acuerdo a la siguiente función:

$$y = 0{,}0026x6 - 0{,}2115x5 + 6{,}6352x4 - 98{,}616x3 + 683{,}09x2 - 1599{,}1x + 1584 \quad (4)$$

Donde y es función polinomial de grado 6, la cual es una línea de tendencia realizada de acuerdo a la gráfica real dibujada para valores de luminosidad reales tomados por hora de un día de otoño. Así para los valores tomados de acuerdo a la Tabla 3, tendremos una gráfica y una línea de tendencia como se ve en la Fig. 2.

Tabla 3. Valores de Luminosidad.

Hora	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00
Luminosidad	500	500	1000	1000	2000	2000	3000	3000
Hora	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00
Luminosidad	3000	3000	3000	3000	3000	3000	3000	3000
Hora	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Luminosidad	3000	3000	3000	3000	3000	3000	1000	1000

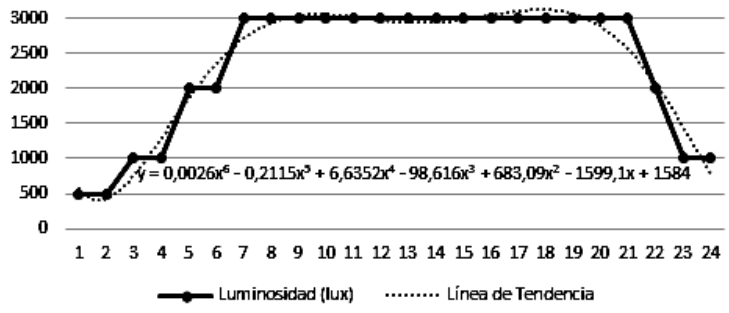


Fig. 2. Gráfico de Función de Tendencia de Luminosidad de un día de otoño en base a los valores sensados reales.

De la misma manera se simulan las otras 3 variables de contexto y uso de actuadores, cuyo uso afecta las condiciones climáticas dentro del invernadero.

3.4 Condiciones óptimas generales para el crecimiento de la planta

Las condiciones óptimas generales para el mejor crecimiento de las plantas requieren una luminosidad mayor, lo cual produce un aumento de la temperatura, la humedad relativa y del dióxido de carbono.
Sin embargo, el aumento de la temperatura disminuye la humedad relativa, lo cual hace necesario algún mecanismo que compense estas condiciones de manera tal que puedan producirse los valores óptimos, es decir, mayor luminosidad, la cual se presenta naturalmente por la radiación solar, que se considera en amplitudes de entre 12 y 16 hs dependiendo de la estación del año y la zona donde se implanta el invernadero, temperatura alta entre 10ºC y 20ºC, valores de humedad relativa alta, concentraciones de dióxido de carbono de entre 0.1 y 0.2%, que se establece en la Tabla 4. Sin embargo, cabe destacar que estas son condiciones generales, ya que cada planta de una especie en particular posee condiciones óptimas propias.

Tabla 4. Condiciones óptimas generales para el proceso de fotosíntesis

Valores sensados	Valores óptimos para la fotosíntesis	Actuadores
Iluminación → Mayor	Mayor	Prender Luminaria cuando oscurezca.
Temperatura → Alta	Alta	Verificar si la temperatura no supera los 20ºC. El rango permitido es de [10ºC,20ºC].
HR → Normal	Alta	Disminuir la temperatura dentro del rango permitido [10ºC, 20ºC], para que la humedad relativa aumente.
CO2 → Normal	Alta	Aumentar la temperatura entre [18ºC, 23ºC] para que el CO2 aumente.

4 Características de la Implementación

Para realizar la implementación del caso de estudio utilizando el modelo MASCO, fue necesario adaptarlo a las características mencionadas en el apartado 3. Se trabajó capa por capa y se recurrió al uso de patrones para dar soporte al proceso de decisión sobre invocación de servicios para corregir valores anómalos en base a las condiciones actuales del contexto, es decir, de las variables de contexto analizadas como un conjunto dada la dependencia que existe entre ellas mencionada en el apartado 3.1, buscando además que exista bajo acoplamiento en todo el procedimiento de monitoreo, control, verificación, decisión y ejecución de servicios de corrección. A continuación se especifican los patrones utilizados en cada una de las capas adaptadas de MASCO al caso de estudio brindando soluciones a los problemas que se presentaron en cada caso.

4.1 Hardware Abstraction Layer

Se especifican las clases que representan los sensores para monitorear las cuatro variables de contexto: Luminosidad, Dióxido de Carbono, Humedad Relativa y Temperatura, y los actuadores necesarios para los dispositivos que las regulan aumentando o disminuyendo sus valores, como se muestra en la Fig. 3.

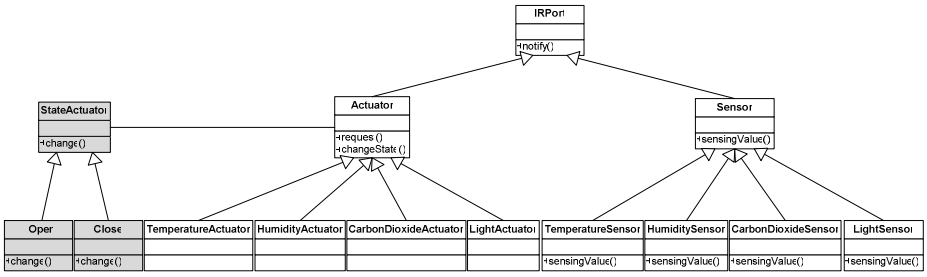


Fig. 3. Hardware Abstraction Layer

Se implementó el patrón State en la clase StateActuator, áreas sombreadas en la Fig. 3, para determinar el estado en el que se encuentran cada uno de los actuadores. Así se establece su disponibilidad para llamar los servicios sobre estos y realizar acciones de regulación sobre los dispositivos como ventanas, riego por goteo, etc. Se decidió utilizar este patrón ya que simplifica la determinación del estado sobre los actuadores. De otro modo significaría colocar sensores a los actuadores agregando complejidad y mezclando los dispositivos de toma de valores de variables de contexto con aquellos que los regulan.

Las clases Open y Close determinan los dos posibles valores de los actuadores, esto indica si el dispositivo regulador sobre el que actúa está actualmente abierto o cerrado, permitiendo llamar a funciones de corrección en las clases que corresponden a los actuadores. Además se especificó una herencia sobre la clase sensor, que permite agregar tipos diferentes de sensores.

4.2 Sensing Concern Layer

Se especifican las políticas para convertir los datos obtenidos por los sensores a datos que pueden ser procesados y entendidos para la toma de decisiones posterior (Fig. 4). La clase SensingConcern recibe los valores de los sensores mediante la implementación de un patrón Observer, que viene determinado desde el Modelo MASCO. SensingConcern no conoce el momento en que son monitoreados los valores de las variables de contexto, Hardware Abstraction Layer envía la notificación en cuanto los valores son monitoreados, esto se realiza en el mismo momento para las cuatro variables de contexto para garantizar la evaluación conjunta de la condición del ambiente. Además, para poder aplicar una política de conversión adecuada de los valores a cada variable de contexto monitoreada se aplica un patrón Strategy, así la clase SensingPolicy determina las reglas de transformación.

Una vez realizada la transformación de los valores se notifica a la clase ContextAggregator de Context Layer.

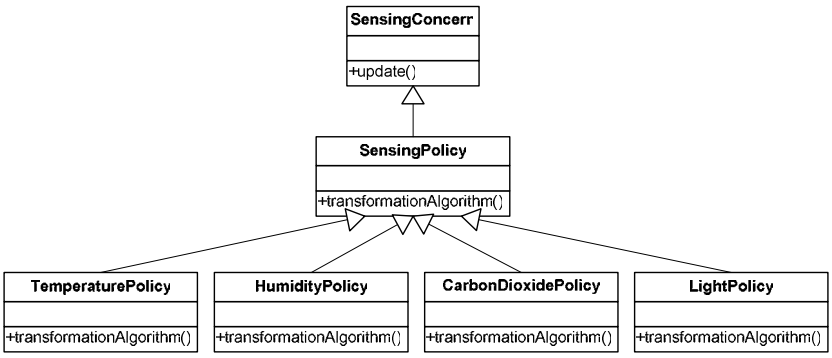


Fig. 4. Sensing Concern Layer

4.3 Context Layer

Esta capa realiza la decisión sobre cuál es la variable de contexto sobre la que se invocará la petición de los servicios, dejando la decisión de cuáles servicios llamar a la capa Service. Para poder llevar a cabo esta decisión se implementaron los patrones State y Template Method de manera combinada, además de los patrones que el modelo MASCO implementa para la clase ContextAggregator: Observer para recibir la notificación cuando sean tranformados los valores monitoreados de las variables de contexto desde la capa Sensing Concern, y Mediator para combinar los valores de las variables de contexto debido a su dependencia.

En esta capa se verifica individualmente si los valores de cada una de las variables de contexto se encuentran en sus valores de referencia, mediante el patrón Mediator se combinan las variables y se verifica el contexto completo a través del patrón Template Method que se implementa en la clase ContextShapeDefinition, esto es así porque es posible que las variables de contexto que se encuentren con valores anómalos puedan presentar un contexto completo anómalo correspondiente al caso más crítico, también un contexto único corresponde a una sola variable de contexto con valores anómalos, mientras que el contexto opuesto se da cuando las variables Temperatura y Humedad Relativa, las que son inversamente proporcionales, poseen valores anómalos y su corrección implica que debe encontrarse un equilibrio entre dos valores opuestos, de acuerdo a lo planteado en el apartado 3.1. Para poder determinar si una variable de contexto posee un valor anómalo se utilizó el patrón State, cambiando su estado luego de la comprobación independiente, y sólo en caso de valores anómalos se notifica a la clase ContextAggregator para que defina la forma del contexto y llame los servicios de acuerdo al caso del contexto anómalo actual. Si no se presenta ningún valor anómalo no se llaman servicios y se aguarda al siguiente tiempo de monitoreo para verificar el estado del sistema. Los patrones State y Template Method añadidos se encuentran sombreados en la Fig. 5.

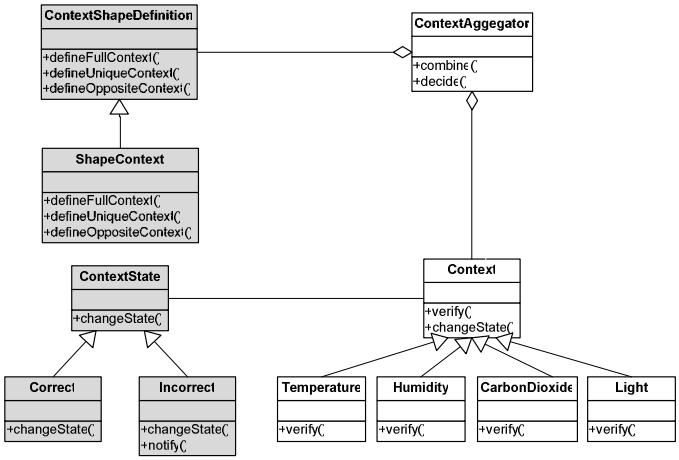


Fig. 5. Context Layer

4.4 Service Layer

En esta capa se produce el proceso de decisión sobre qué servicios deben solicitarse en base a la decisión tomada en la capa Context. Esto es en base a la variable de contexto con mayor prioridad que posee valores anómalos, ya que la relación de dependencia garantiza que alterar una variable de contexto altera las demás de acuerdo a la relación planteada en el apartado 3.1. Se implementó el patrón Strategy, marcado en gris en la Fig. 6 para realizar el llamado específico de los servicios de aumentar o disminuir los valores de esa variable de contexto. Para esto la clase ServiceCoodinator realiza la decisión en base al estado de los Actuadores, como se describió en el apartado 4.1.

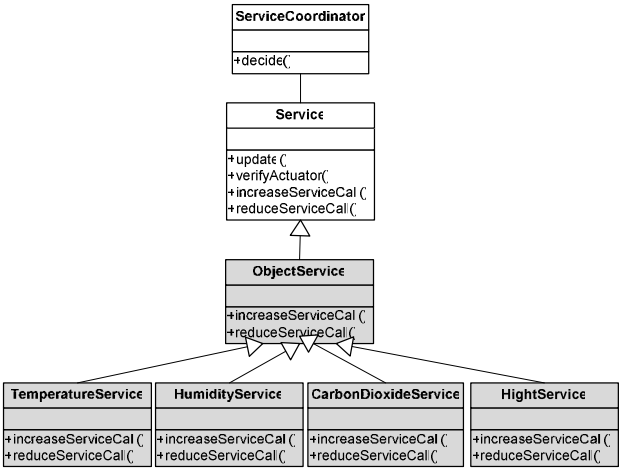


Fig. 6. Service Layer

Una vez llamados los servicios se cambia el estado de los Actuadores en la capa HardwareAbstraction y se aguarda al siguiente período de monitoreo donde se valida la acción realizada. De ocurrir un estado crítico de contexto completo no corregido en tres iteraciones se pasa al control manual, porque hay riesgo de pérdida de cultivo por condición extrema que no puede corregirse de forma automatizada.

5 Conclusiones

La aplicación de patrones al modelo brindó soluciones a dos grandes problemas que se presentaron al implementar el caso de estudio: la interacción entre variables con dependencia y la decisión sobre cuales servicios solicitar para corregir valores anómalos. Su uso garantizó la integridad estructural del modelo, permitiendo adaptar las capas incorporando las clases necesarias, mostrando MASCO flexibilidad al aplicarlo al caso de estudio

El rendimiento, que podría haber sido un inconveniente al presentarse la solicitud de servicios que van encadenados a una capa inferior desde la aplicación, fue solucionado brindando comunicación directa entre capas, como es el caso de los servicios solicitados desde la Context Layer a Service Layer, manteniendo las características propias de un modelo en capas.

6…Referencias

1. Gordillo, S., Rossi, G., Fortier, A.: Engineering Pervasive Services for Legacy Software. Proccedings of the 1st International Workshop on Software Engineering for Pervasive Services, Lyon (2006)
2. Dey, A. K.: Providing architectural suport for building context aware applications. PHD Thesis. Georgia Institute Technology, USA (2001)
3 García, M.: Automatización de Procesos Industriales. Universidad Politécnica de Valencia, España (1999)
4. Quincoces, V.E., Gálvez, M.P., Cáceres, N.R., Vega, A.A., Ramos, H. O.: Extensión de un modelo en capas que provee servicios para aplicaciones sensibles al contexto. Investigaciones en Facultades de Ingeniería del NOA, ISBN 978-987-633-041-1.2009, Vol I, pp. 35--40, Cap. IV. EUNSa, Salta (2009)
5 Quincoces, V.E., Gálvez, M.P., Cáceres, N.R., Vega, A. A.: Modelo que provee servicios para aplicaciones sensibles al contexto: Validación en etapas tempranas. Investigaciones en Facultades de Ingeniería del NOA, pp. 481-486, EdiUNJu, Argentina (2010)
6 Gálvez, M.P., Quincoces, V.E., Cáceres, N.R., Vega, A.A.: Refinamiento de un Modelo en Capas que Provee Servicios de Ubicación para Aplicaciones Sensibles al Contexto. III Congreso Internacional de Telecomunicaciones, Tecnologías de la Información y las Comunicaciones, Quito (2010)
7 Gálvez, M.P., Brouchy,C., González, O., Cáceres, N.R., Quincoces, V. E.: Modelo que provee servicios para aplicaciones sensibles al contexto (MASCO): Interacción entre entidades. Investigaciones en Facultades de Ingeniería del NOA, pp. 1103-1109. Científica Universitaria, UNCa, Argentina (2011)

Um mecanismo de captura de informações contextuais em um Ambiente de Desenvolvimento Distribuído de Software

Yoji Massago[1], Renato Balancieri[1,] Raqueline Ritter de Moura Penteado[1], Elisa Hatsue Moriya Huzita[1,] Rafael Leonardo Vivian[2]

1. Universidade Estadual de Maringá – UEM, Maringá, Paraná, Brasil
yojimassago@gmail.com, {rbalancieri2, rrmpenteado, ehmhuzita}@uem.br
[2].Instituto Federal do Rio Grande do Sul – IFRS, Sertão, Rio Grande do Sul, Brasil
rafavivian@gmail.com

Abstract. The Distributed Software Development (DSD) has been increasingly adopted by software development companies since it provides support for a better use of their material, human and time resources. However, it presents challenges from geographical distance, time and cultural differences. So, different teams are faced to communication problems which affect directly the quality of the product generated. During software development several artifacts are generated, and they may change during their life. Therefore it is important that such information can be captured and properly dealt in order to be disseminated and used. The purpose of this paper is to present a mechanism to capture contextual information of Java source code files, from a repository for distributed version control. Then, they are stored in a XML file, making they more flexible to be used by other tools. It will also contribute to improve production in DSD.

Keywords: Contextual Information, Data Repository, Distributed Software Development.

1 Introdução

Em busca de vantagens competitivas e cooperativas, diversas organizações adotaram atividades multilocais, multiculturais e globalmente distribuídas, aumentando a produtividade, melhorando a qualidade e reduzindo custos de suas tarefas ([11][5][1]).

O Desenvolvimento Distribuído de Software (DDS) surgiu para tentar resolver vários problemas que, frequentemente, existiam no desenvolvimento de software tradicional, principalmente no que se refere a alocação de recursos e o aproveitamento do tempo. Entretanto, a dispersão geográfica, a distância temporal e as diferenças sociais, elementos inerentes a DDS, ampliaram alguns dos desafios existentes no desenvolvimento de software e, principalmente, adicionaram novas exigências acerca da comunicação entre os indivíduos participantes de um trabalho cooperativo [11].

Um dos desafios é a captura e a disseminação de informações sobre o processo de desenvolvimento entre os integrantes de uma equipe.

Para tentar resolver o problema da captura e disseminação de informação contextual, uma solução seria fazer uma constante verificação dos dados e, então, capturar informações consideradas importantes para serem repassadas aos membros de uma equipe com o intuito de aumentar a compreensão do contexto. Porém, se os dados capturados forem armazenados de forma não padronizada, o uso posterior dos mesmos pode se tornar inviável. Logo, é necessário algum meio para capturar as informações relevantes e armazená-las em um formato que pode ser usado como padrão. Esta padronização pode ser alcançada utilizando-se por exemplo arquivos no formato XML *(Extensible Markup Language)* e XMI (XML *Metadata Interchange*).

Durante o desenvolvimento de um software são criados muitos artefatos, tais como documento de requisitos, diagramas, códigos, relatórios, entre outros. Uma abordagem para apoiar a percepção sobre a criação, a evolução e a manutenção do código fonte, dos diagramas de classes e dos relatórios de erros que contêm justificativas de mudanças efetuadas nestes artefatos, foi proposta em [11]. Nesta abordagem, estes artefatos são armazenados em um repositório de dados central, sendo os relatórios armazenados em um banco de dados, os diagramas em formato XMI, e os códigos fonte em formato texto.

Os arquivos de código fonte, normalmente, estão armazenados em formato tipo texto de acordo com a gramática da linguagem de programação, adotada para o desenvolvimento de uma determinada aplicação. Dependendo da gramática e do tamanho do código, para se obter informações do mesmo pode demandar muito tempo e poder computacional. No caso destas mesmas informações serem usadas mais de uma vez, como nas informações necessárias a todos os membros de uma equipe distribuída, refazer tudo, a cada vez que a informação for necessária, pode ser muito custoso.

Assim, este artigo tem por objetivo apresentar um mecanismo que capture informações a partir de arquivos de código fonte Java armazenados em um repositório de Sistema de Versionamento Distribuído e armazenados em formato XML, para possibilitar maior facilidade de manipulação destas informações por parte de outros sistemas/programas. Este mecanismo será útil para os casos em que existem muitos dados nos arquivos de código fonte Java, dos quais se necessita capturar apenas uma parte das informações contidas para serem utilizadas por outros programas, e, também, serem acessadas e manipuladas facilmente por uma equipe de DDS.

O texto encontra-se dividido em mais quatro seções além da introdução. A seção 2 descreve os conceitos relacionados ao desenvolvimento do mecanismo de captura de informação. A seção 3 mostra como foi desenvolvido do mecanismo: os passos executados, as funcionalidades desenvolvidas. A seção 4 apresenta os resultados obtidos, bem como a avaliação destes. Por último, a seção 5 apresenta as conclusões.

2 Conceitos envolvidos no projeto do mecanismo

Durante muitos anos, o desenvolvimento de software ocorria de forma centralizada, por uma equipe local. Mas, dependendo das empresas e do software a ser

desenvolvido, trabalhar utilizando apenas os recursos locais (humanos e/ou materiais) era difícil e custoso. Assim, buscando melhorar o desenvolvimento, otimizar a alocação de recursos, diminuir o custo, entre outros fatores, surgiu a ideia de DDS.

"*O DDS tem sido caracterizado principalmente pela colaboração e cooperação entre departamentos de organizações e pela criação de grupos de pessoas que trabalham em conjunto, mas estão localizados em cidades ou países diferentes, distantes temporal e fisicamente.*"[8]. Resumidamente, DDS é o desenvolvimento de um software por uma equipe cujos membros estão trabalhando em locais geograficamente dispersos.

"*Para a execução de um trabalho colaborativo por um grupo de pessoas, é importante que os indivíduos compreendam as atividades dos outros para que isso seja relevante para a realização de suas próprias tarefas e, assim, otimizar o andamento dos trabalhos. Percepção ou Awareness, é uma compreensão das atividades dos outros, que oferece um contexto para a própria atividade do indivíduo.*" [4] apud [11].

Contexto é qualquer informação que pode ser usada para caracterizar uma situação de uma entidade. Uma entidade é uma pessoa, um lugar, ou um objeto que é considerado relevante para a interação entre um usuário e uma aplicação, incluindo o próprio usuário e a própria aplicação [3] apud [11].

No desenvolvimento de software de forma colaborativa, faz-se necessário a percepção dos eventos que estão ocorrendo, a fim de melhorar o desempenho do desenvolvimento, seja evitando trabalhos repetidos/duplicados, seja melhorando na compreensão conjunta da equipe, ou através de outros fatores. No DDS, isso se torna mais importante, considerando-se os desafios decorrentes desta abordagem de desenvolvimento (dispersão geográfica, diferença de fuso horário,...).

Para evitar perdas de dados importantes, muitos desenvolvedores de software utilizam algum Sistema de Controle de Versões (SVM) para o armazenamento de seus dados. Um SVM "*... é responsável por manter o histórico de diversas revisões de uma mesma unidade de informação. Ele é comumente utilizado em desenvolvimento de software para gerenciar o código fonte de um projeto*."[2].

Uma das variações dos SVM é o Sistema de Versões Concorrentes (CVS – *Concurrent Version System*), o qual permite o armazenamento dos dados em um repositório local ou remoto. Neste repositório são armazenadas as versões antigas e os logs daqueles que manipularam os arquivos e quando isto foi feito. Como exemplos de CVS podem ser citados o Subversion[1], o Perforce[2], entre outros.

Porém, um CVS comum, principalmente pelo fato de utilizar um repositório central, não possui suporte adequado para a sua utilização no Desenvolvimento Distribuído de Software [6]. Assim surgiram os DCVS *(Distributed Concurrent Versions System)* que fornecem o suporte adequado para que os programadores que estão em locais geograficamente dispersos possam desenvolver software colaborativamente. Estes sistemas utilizam vários repositórios espalhados pela rede, e também possuem um mecanismo com o qual os repositórios filhos possam

1 http://subversion.apache.org/

2 http://www.perforce.com/products/perforce

compartilhar informações entre si de modo paralelo [6]. Alguns dos DCVS existentes atualmente são o Git[3], e o Mercurial[4].

Dentre estes, para o desenvolvimento do mecanismo apresentado neste artigo foi utilizado DCVS Mercurial, que é uma ferramenta multiplataforma que pode ser utilizada em conjunto com algumas IDEs *(Integrated Development Environment)* como NetBeans[5] e Eclipse[6]. Algumas outras características desta ferramenta são: é um programa de linha de comando, possui repositório descentralizado, é escalável, suporta trabalho distribuído e colaborativo, e suporta arquivos texto e binário.

3 Detalhes do Projeto do Mecanismo

Uma abordagem para a percepção de informações contextuais sobre os artefatos de software no ambiente de desenvolvimento distribuído de software chamado DiSEN é proposta em [11]. Em [12] é apresentada a abordagem, da qual faz parte um mecanismo para exibir as relações existentes entre os arquivos de código fonte (em Java) e os diagramas de classe correspondente. Além destas relações foi prevista em [12] a necessidade de oferecer apoio adequado à captura de informações em arquivos de código fonte Java que pudessem proporcionar percepção de contexto aos membros de equipes DDS.

Assim, o mecanismo apresentado no presente artigo faz parte do trabalho apresentado em [12] e tem como foco obter informações contextuais a partir de arquivos de código fonte escritos em Java e armazenados no repositório Mercurial. Para tal, faz se necessário verificar a localização destes códigos, para, posteriormente, realizar uma varredura através destes a fim de conseguir as informações desejadas, além de manipular os métodos/comandos existentes no sistema Mercurial para a manipulação/obtenção dos dados contidos no repositório, tais como versões anteriores, quem executou estas mudanças e datas e horários das modificações.

3.1 Desenvolvimento

As principais funcionalidades do mecanismo ora apresentado são: (i) capturar informações do sistema Mercurial, (ii) verificar os programas Java existentes e, após as verificações, (iii) criar os arquivos XML contendo as informações capturadas.

Assim, o desenvolvimento do mecanismo, obedeceu-se as seguintes etapas:

1. *Captura de informação no Mercurial* - Para isso, foi criado um arquivo de *shell script* que executa os comando do Mercurial e chama o mecanismo criado, passando os dados relevantes como parâmetros de entrada. Ele irá capturar informações sobre os arquivos modificados, a versão dos arquivos, o autor do *commit*, a data da modificação, a mensagem enviada junto ao *commit* e a versão do Mercurial;

3 http://git-scm.com/
4 http://mercurial.selenic.com/
5 http://netbeans.org/
6 http://www.eclipse.org/

2. *Verificação dos arquivos Java existentes* - Para esta etapa, foram criadas funções que fazem a varredura de um diretório, previamente especificado pelo usuário do sistema (no caso, o diretório do repositório Mercurial), bem como os seus subdiretórios, em busca de arquivos de código fonte Java;
3. *Verificação dos arquivos de código fonte Java* - Para a varredura dos arquivos de código fonte Java, foram implementadas funções baseadas em um compilador: mais especificamente, as funcionalidades de análise léxica e sintática, responsáveis por determinar se os arquivos estão de acordo com a gramática específica da linguagem de programação. Simultaneamente, é utilizado a biblioteca Java *jdom-2.0.1*[7], para manipular arquivos XML, a fim de armazenar as informações obtidas;
4. *Execução junto ao commit* - Também foi modificado um dos arquivos do Mercurial (o arquivo *hgrc*), para que execute o *shell script* acima mencionado, toda vez que um *commit* é executado.

3.2 Visão geral do projeto

Para facilitar a compreensão, são apresentados a seguir dois diagramas do projeto do mecanismo. A Fig. 1 exibe os pacotes e as classes existentes, bem como as relações destas classes. Nela pode-se observar que existem três pacotes:

- Pacote *criarxml*, que contém a classe CriacaoArquivoXML, responsável pela criação e armazenamento de dados em arquivos XML;
- Pacote *verificarepositório*, que contêm as classes *main*, Parametros e ManipulacaoDiretorio. A classe Parametros é um objeto que armazena os dados que a classe main recebeu como entrada (os dados obtidos pelo shell script), a fim de serem repassados ao Parser, onde serão armazenados nos arquivos XML. A classe ManipulacaoDiretorio é responsável pela varredura dos diretórios em busca de arquivos .java.
- Pacote *comp*, que contêm as partes do "compilador": Lexer e Parser, além do objeto Token. O Lexer é o analisador léxico, responsável por analisar o arquivo e criar tokens, com os dados necessários. No Parser ou analisador sintático, que existe a "varredura" das informações contidas nos arquivos Java, bem como o seu armazenamento em formato XML, utilizando-se a classe CriacaoArquivoXML.

A Fig. 2 mostra um diagrama de sequência do funcionamento geral do mecanismo proposto. Nela pode-se observar que a classe *main* (VerificacaoRepositorio) inicializa uma nova instância da classe ManipulacaoDiretorio e inicializa-o por meio do método ImprimirConteudo. Nesta classe ManipulacaoDiretorio existe um *loop* que fica ativo até que não exista mais arquivo .java a ser verificado. Dentro deste *loop*, sempre que um arquivo Java é encontrado, é criado um novo Parser e inicializa-o. Este Parser utiliza o Lexer para capturar os dados (*Tokens*) dos arquivos e armazena os dados em XML, com a utilização da classe CriacaoArquivoXML.

7 http://www.jdom.org/

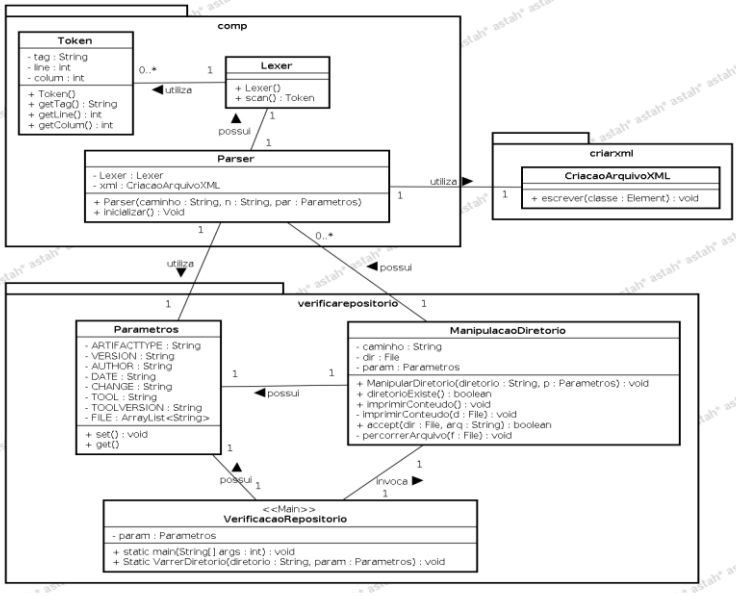


Fig. 1 Diagrama de pacotes do mecanismo

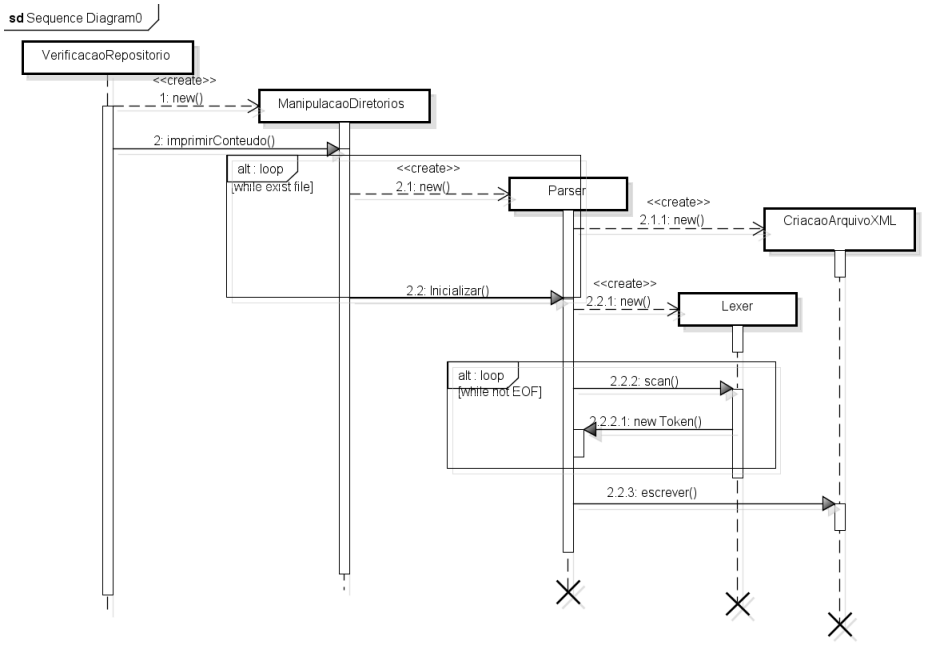


Fig. 2 Diagrama de sequência mostrando o fluxo de execução do mecanismo

4 Avaliação dos resultados

Para a validação do mecanismo desenvolvido, foram definidos casos de teste, para mostrar exemplos de uso deste. Para tanto, executa-se o mecanismo criado, com diversas entradas, e observam-se as saídas a fim de verificar se estas estão de acordo com o esperado. Este método de avaliação foi escolhido para verificar a corretude do mecanismo.

Para os testes, foram escolhidos vários casos, entre eles: um código simples, um código mais complexo, comparação dos XML após algumas modificações. Destes, será apresentado o caso de um código simples, sendo os outros possíveis de serem visualizados em [7].

Para exemplificar a geração de um arquivo XML a partir de uma classe Java, a seguir será mostrado o caso de um código simples. A Fig. 3 mostra o arquivo texto de um código fonte Java de um arquivo chamado *PersistenceAcessoPolicy.java*. Conforme pode ser observado na linha 5, este arquivo está contido no pacote *disen.supernode*. Pelas linhas 7 a 11, pode-se observar que utiliza cinco *imports*, dos quais três são do próprio DiSEN, e os outros dois são bibliotecas do Java. Outros dois pontos a serem considerados são: este arquivo é uma classe (linha 17), com o nome *PersistenceAcessoPolicy* e ela possui um único método identificado como *validarUsuario* (linha 20).

```
1  /*
2   * To change this template, choose Tools | Templates
3   * and open the template in the editor.
4   */
5  package disen.supernode;                                          <- Pacote
6
7  import br.uem.din.disen.core.comunicacao.acesso.AcessoPolicy;
8  import br.uem.din.disen.core.comunicacao.acesso.TipoAcesso;
9  import disen.recurso.usuario.bean.Usuario;                        Importações
10 import java.util.logging.Level;
11 import java.util.logging.Logger;
12
13 /**
14  *
15  * @author will
16  */
17 public class PersistenceAcessoPolicy implements AcessoPolicy {    <- Classe
18
19     @Override
20     public TipoAcesso validarUsuario(String login, String senha) { <- Método
21         try {
22             Usuario u = new Usuario();
23             u = u.getUsuarioByLogin(login);
24             if(u!=null) {
25                 if(u.getSenha().equals(senha)) {
26                     return TipoAcesso.NODE;
27                 }
28             }
29         } catch (Exception ex) {
30             Logger.getLogger(PersistenceAcessoPolicy.class.getName()).log(Level.SEVERE, null, ex);
31         }
32         return TipoAcesso.NEGADO;
33     }
34 }
```

Fig. 3 Arquivo de código fonte java

A Fig. 4 mostra parte do arquivo XML gerado pelo mecanismo, que é um arquivo de nome *PersistenceAcessoPolicy.xml*, que possui os campos *artifacttype*, *version*,

author, *date*, *change*, *tool*, *toolversion*, *package*, *Imports* e *type*. A primeira *tag* se refere ao tipo de artefato armazenado (no caso um arquivo de código fonte). Os seis atributos seguintes são referentes aos dados obtidos do Mercurial: i) versão do artefato, ii) autor do *commit*, iii) a data da execução do *commit*, iv) a mensagem de *commit*, v) o nome do repositório e vi) a versão deste repositório. Abaixo destas *tags*, existem as *tags package*, *Imports* e *type*. Destes, a primeira se refere ao Pacote no qual o arquivo Java está inserido; a segunda é referente aos *imports* utilizados e, finalmente, o *type* (no caso, *classe*), no qual os dados da classe estão inseridos.

```
- <PersistenceAcessoPolicy>
    <artifacttype>SourceCode</artifacttype>
    <version>51</version>
    <author>Yoji Massago <massagoy@gmail.com></author>
    <date>2012-11-05 20:50 -0200</date>
    <change>teste</change>
    <tool>Mercurial</tool>
    <toolversion>Mercurial Distributed SCM (version 2.0.2)</toolversion>
    <package>disen.supernode</package>
  - <Imports>
    - <import1>
        br.uem.din.disen.core.comunicacao.acesso.AcessoPolicy
      </import1>
    - <import2>
        br.uem.din.disen.core.comunicacao.acesso.TipoAcesso
      </import2>
      <import3>disen.recurso.usuario.bean.Usuario</import3>
      <import4>java.util.logging.Level</import4>
      <import5>java.util.logging.Logger</import5>
    </Imports>
  - <type>
      <type>classe</type>
    - <classe>
        <modifier>public</modifier>
        <NomeClasse>PersistenceAcessoPolicy</NomeClasse>
        <implements>AcessoPolicy</implements>
      - <method>
          <modifier>public</modifier>
          <retorno>TipoAcesso</retorno>
          <nome>validarUsuario</nome>
        - <parametros>
          - <parametro>
              <tipo>String</tipo>
              <identificador>login</identificador>
            </parametro>
          - <parametro>
              <tipo>String</tipo>
              <identificador>senha</identificador>
            </parametro>
          </parametros>
```

Fig. 4 Parte do XML gerado para o arquivo PersistenceAcessoPolicy.java

Como resultado dos testes foi verificado que os arquivos que foram testados e verificados (o que inclui uma versão antiga do DiSEN, além do próprio mecanismo),

retornam os resultados esperados, capturando as informações e armazenando corretamente em arquivos XML.

Devido ao grande número de arquivos gerados ao utilizar este mecanismo nos arquivos de código fonte Java do DiSEN, não foram verificados todos os arquivos, um por um, para garantir que todos foram varridos e verificados de forma correta. Todavia, aqueles que foram verificados estavam de acordo com o esperado, da mesma forma que os exemplos mostrados anteriormente.

5 Conclusão

Este artigo apresenta um mecanismo para captura de informações contextuais a partir de um repositório de controle de versão, mais especificamente do Mercurial. Para tal, realizou-se, inicialmente, uma pesquisa para entender os conceitos relevantes, a serem utilizados, e a partir de tais informações, optou-se pelo uso dos analisadores léxicos e sintáticos de um compilador.

O analisador léxico-sintático é um mecanismo para verificar se um arquivo de código fonte possui a sintaxe correta. Também, existem vários comandos no Mercurial, que podem ser utilizados, como no caso deste mecanismo, em conjunto, por meio de um *script*, para obter informações, solicitar alguma alteração, entre outros tipos de comandos.

Além dos testes apresentados na seção 4 deste artigo foram também realizados outros que por limitações de espaço não estão aqui ilustrados. Detalhes de tais testes podem ser encontrados em [7].

Ao final da implementação, e tendo como base os testes realizados, verificou-se que este mecanismo consegue obter as informações necessárias para o nosso contexto. No caso desta versão do mecanismo, adotou-se a estratégia de capturar a maioria das informações do código fonte e colocá-las nos arquivos XML. Isso foi feito pelo simples fato de que, como se utiliza os analisadores, será percorrido o arquivo inteiro e, consequentemente, obter uma pequena parte ou uma grande parte do código não faria muita diferença, além de que, no dado momento, não se sabe quais as informações que outras ferramentas irão necessitar. Futuramente, no caso de não necessitar destes dados "extras", basta retirar as linhas do código referentes à inserção, destes mesmos dados, no arquivo XML.

Assim, espera-se que este mecanismo possa ajudar os desenvolvedores e principalmente outras ferramentas, como o mecanismo desenvolvido por [11], na captura de informações sobre o desenvolvimento dos arquivos de código fonte escritos em Java e, consequentemente, possa ser útil para que outros mecanismos e/ou pessoas consigam obter as informações de contexto necessárias ao bom desempenho das equipes geograficamente distribuídas.

Uma das limitações existentes neste mecanismo está no uso dos analisadores, o que implica no fato deste mecanismo ser específico à gramática escolhida na criação destes analisadores. Assim, no caso de utilizar para outras linguagens ou versões, serão necessários ajustes nestes analisadores.

Portanto, um possível trabalho futuro será modificar os analisadores a fim de poder reconhecer outras gramáticas de outras versões ou até de outras linguagens. Também,

seria interessante a existência de alguma funcionalidade que permitisse o mecanismo identificar, de forma automática, a linguagem utilizada nos arquivos a serem verificados. Com a implementação destes dois casos, o mecanismo poderá fazer a varredura e transformação em XML de arquivos de várias Linguagens de Programação, de forma automática e sem a necessidade do usuário informar, a cada vez, a linguagem utilizada, ou utilizar apenas uma linguagem específica para programar o sistema inteiro. Outro trabalho a ser desenvolvido no futuro será a integração deste mecanismo com o projetado pelo [11][12] com intuito de melhorar o desempenho deste último.

Referencias

1. AUDY, J.: PRIKLADNICKI, R.: Desenvolvimento Distribuído de Software: desenvolvimento de software com equipes distribuídas. Rio de Janeiro: Elsevier, (2008)
2. AUVRAY, S.: Melhores da InfoQ em 07: Sistemas de Controle de Versão Distribuído: Um Guia não tão rápido. InfoQ, (2007)
3. DEY, A. K.: Providing Architectural Support for Building Context-Aware Applications. Ph.D. Thesis, Georgia Institute of Technology, (2000)
4. DOURISH, P., BELLOTTI, V.: Awareness and Coordination in Shared Workspaces. In: The 1992 ACM Conference on Computer-Supported Cooperative Work, *Proceedings*…, pp. 107--114, (1992)
5. HUZITA, E.H.M., SILVA, C.A., WIESE, I.S., TAIT, T.F.C., QUINAIA, M., SCHIAVONE, F.L.: Um conjunto de soluções para apoiar o desenvolvimento distribuído de software. In: II Workshop de Desenvolvimento Distribuído de Software, Campinas, pp. 101--110, (2008)
6. GIT.: Reference Manual. Disponível em: <http://git-scm.com/documentation>. Acesso em: 11/09/2013, (2013)
7. MASSAGO, Y.: Um mecanismo para captura de informações contextuais em um ambiente de desenvolvimento distribuído de software, Departamento de Informática, Universidade Estadual de Maringá, Trabalho de Conclusão de Curso, (2012)
8. PRIKLADNICKI, R., LOPES, L., AUDY, J. L. N., EVARISTO, R.: Desenvolvimento Distribuído de Software: um Modelo de Classificação dos Níveis de Dispersão dos Stakeholders. University of Illinois at Chicago - College of Business Administration 601 S. Morgan Street MC 294, Chicago, IL 60607, United States, (2004)
9. VIEIRA, V.: Gerenciamento de contexto em sistemas colaborativos. Tese de Doutoramento, CIn - Universidade Federal de Pernambuco, Recife - Pernambuco, Monografia de Qualificação, (2006)
10. VIEIRA, V.: CEManTIKA: A Domain-Independent Framework for Designing Context- Sensitive System. Tese de Doutorado, Recife: Cin – UFPE, (2008)
11. VIVIAN, R. L.: Uma Abordagem Context-Awareness sobre Artefatos de Software no Desenvolvimento Distribuído de Software. 29f. Projeto de Dissertação (Mestrado) - Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual de Maringá, Maringá, (2011)
12. VIVIAN, R. L., HUZITA, E. H. M., LEAL., G. C. L.: Supporting distributed software development through context awareness on software artifacts: the DiSEN-CollaborAR approach. In: 28th Annual ACM Symposium on Applied Computing *(SAC '13), Proceedings…*, New York, NY, USA, pp. 765-770, (2013)

Q-Scrum: una fusión de Scrum y el estándar ISO/IEC 29110

Ariel Pasini[1], Silvia Esponda[1], Marcos Boracchia[1,] Patricia Pesado[1, 2]

[1]Instituto de Investigación en Informática LIDI (III-LIDI),
Facultad de Informática, UNLP, 50 y 120, La Plata, Buenos Aires, Argentina
[2] CIC (Comisión de Investigaciones Científicas de la Provincia de Bs. As.), Argentina

{apasini, sesponda,marcosb,ppesado}@lidi.info.unlp.edu.ar

Abstract. Se realiza una comparación de la metodología de desarrollo ágil Scrum con los requerimientos del estándar ISO/IEC 29110. Se analizan las competencias de los roles ISO/IEC 29110 versus los roles de la metodología Scrum, los documentos que exige el estándar ISO/IEC 29110 contra los documentos definidos por Scrum y las actividades definidas en el estándar ISO/IEC 29110 respecto las de Scrum. Se presenta Q-Scrum, el modelo propuesto que permite una aproximación de los desarrollos en Scrum a los requerimientos del estándar ISO/IEC 29110, fusionando los roles, documentos y actividades de ambos modelos.

Keywords: Ingeniería de Software -Calidad – PyMEs - Metodologías ágiles - Scrum – ISO/IEC 29110

1 Introducción

Las metodologías ágiles representan una alternativa para el desarrollo de sistemas de software, centrada en el factor humano y el producto software, valorizando la relación con el cliente y el desarrollo incremental del software. Estas metodologías ofrecen entregas frecuentes de software funcional, permitir cambios de requerimientos y participación directa del cliente en el desarrollo. Una de estas metodologías es Scrum [1], que se define como un proceso iterativo incremental y empírico para administrar y controlar el trabajo de desarrollo. Actualmente Scrum es la metodología más utilizada en PyMEs desarrolladoras de software. La decisión de implantar metodologías de desarrollo, indica que la organización ha adquirido experiencia y se encuentra en un proceso de madurez que es de esperar, se afiance con el tiempo [2][3][4].

El estándar ISO/IEC 29110 Perfil Básico es un conjunto de buenas prácticas en el desarrollo del software para asistir y evaluar a las PyMEs desarrolladoras de software en el proceso de mejora. Está compuesto del Proceso de Administración de Proyecto (AP) y del Proceso de Implementación Software (IS), cada uno de ellos posee un

conjunto de roles, actividades y documentos externos, que se deben satisfacer al momento de evaluar el estado de los procesos [5].

En el camino de obtener mejor calidad en las empresas desarrolladoras de software de pequeño y mediano porte, surge la necesidad de compatibilizar la utilización de metodologías ágiles tipo Scrum y estándares de buenas prácticas como ISO/IEC 29110. Sin embargo la estructura y documentación definida por Scrum para sus desarrollos es insuficiente para satisfacer los requisitos del estándar ISO/IEC 29110, por lo cual es necesario desarrollar un nuevo modelo.

Q-Scrum es una propuesta de modelo orientada a PyMES, que proporciona una estructura de roles, documentos y actividades capaces de satisfacer el estándar, con la idea que las empresas la puedan usar como punto de partida en la mejora de sus procesos de desarrollo.

En la sección 2 y 3 se presenta una breve descripción de Scrum y el estándar ISO/IEC 29110, respectivamente, haciendo hincapié en los roles, documentos y actividades, que luego serán relacionados y comparados en la sección 4. En la 5 se presenta Q-Scrum, como una adaptación de Scrum para satisfacer los requisitos del estándar ISO/IEC 29110. Por último se presentan las conclusiones obtenidas del presente trabajo en la sección 6.

2 Scrum

Scrum es un marco de trabajo ágil para desarrollo de software. El trabajo se organiza en ciclos llamados sprints que son iteraciones de corta duración, típicamente de 2 a 4 semanas. Durante cada sprint, el equipo selecciona un conjunto de requerimientos de una lista priorizada, de manera que las funciones desarrolladas al principio del proyecto son las de más alto valor. Al final de cada sprint se entrega un producto de software ejecutable en el ambiente requerido por el cliente. No es un

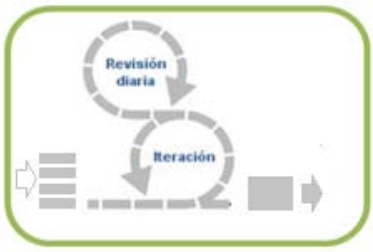


Fig. 1. Scrum

proceso prescriptivo, no describe qué hacer en cada circunstancia, sólo ofrece un marco de trabajo y un conjunto de prácticas que mantienen todo visible y guían los esfuerzos para obtener el resultado más valioso posible.

Scrum coloca todas sus prácticas en un proceso con estructura iterativa e incremental. Esto se muestra en la figura 1, donde el lazo mayor representa una

iteración, que se repite en el tiempo, y que abarca las actividades de desarrollo. La salida de la iteración es un incremento del producto. El lazo más pequeño representa la inspección diaria que tiene lugar durante la iteración, en la cual los miembros del equipo se reúnen para inspeccionar las actividades de todos los miembros y hacer las adaptaciones apropiadas. La iteración es dirigida por la lista de requerimientos (Product backlog). Este ciclo se repite hasta que finalice el proyecto.

Al comienzo de cada iteración, el equipo revisa lo que debe hacer y selecciona lo que cree que se puede convertir en un incremento de la funcionalidad potencialmente entregable. Luego el equipo trabaja haciendo su mejor esfuerzo en el resto de la iteración. Al final de la iteración, el equipo presenta el incremento de la funcionalidad, el cual es construido de manera tal que los involucrados puedan inspeccionar la funcionalidad y oportunamente hacer adaptaciones al proyecto [6] [7],[8].

Las tablas 1 y 2 presentan los roles y documentos que se utilizan en Scrum

Rol	Competencia
Product Owner	El PO representa a quien tiene un interés en el proyecto y el producto resultante. Sus principales responsabilidades son: definir los requerimientos del producto a desarrollar durante el proyecto, ajustar los requerimientos y prioridades a lo largo de todo el proyecto, aceptar o rechazar el producto de software.
Scrum Master	El SM es el líder que facilita el trabajo. Es responsable del proceso de Scrum, de ser necesario enseñándoselo a cada uno de los involucrados en el proyecto. Se asegura de que cada uno sigue las reglas y prácticas de Scrum. Sus principales responsabilidades son conducir la reunión Daily Scrum (DS), conocer el estado de las tareas, identificar barrera y dependencias que impidan el flujo de Scrum y observar y resolver conflictos personales.
Scrum Team	El Equipo es interdisciplinario y con 7±2 integrantes que son los encargados de conocer cómo convertir los requerimientos en un incremento de la funcionalidad y de desarrollar dicho incremento.

Table 1. Roles Scrum

Documentos	Descripción
Product Backlog.	Es un documento de alto nivel para todo el proyecto. Contiene descripciones genéricas de todos los requerimientos, funcionales y no funcionales, contiene estimaciones realizadas a grandes rasgos, tanto del valor para el negocio, como del esfuerzo de desarrollo requerido, la prioridad de las diferentes tareas, etc. Es dinámico, nunca está completo, evoluciona junto con el producto

Sprint Backlog.	Documento detallado que contiene las tareas que el Team va a implementar durante el presente sprint.

Table 2. Documentos Scrum

3 ISO/IEC 29110

La industria de software en PyMEs creció exponencialmente en los últimos años, pero carecía de estándares o modelos de mejora que tuviesen en cuenta la estructura y capacidad interna de las mismas. Con la intención de ayudar a este sector, ISO a través del SC7-WG24, inició su trabajo para lograr que sus estándares de procesos de software (o adaptaciones de éstos) se pudieran aplicar a pequeñas y medianas empresas desarrolladoras de software.

Este grupo estableció un marco común para describir perfiles evaluables del ciclo de vida de software para uso en PyMEs.

La norma define tres perfiles: Perfil Básico, Perfil Intermedio y Perfil Avanzado. El primero de los perfiles se ha publicado en el año 2010 bajo el nombre de ISO/IEC 29110 Perfil Básico, los otros aún permanecen en desarrollo.

El Perfil Básico está compuesto del **Proceso de Administración de Proyecto (AP)**, con el objetivo de establecer y llevar a cabo de manera sistemática las tareas de los proyectos de implementación de software, cumplir con los objetivos del proyecto en calidad, tiempo y costo esperados, y del **Proceso de Implementación Software (IS)**, con el propósito de asegurar la realización sistemática de las actividades de análisis, diseño, construcción, integración y pruebas de productos de software, nuevos o modificados de acuerdo a los requisitos especificados [5].

Cada uno de estos procesos, está compuesto por un conjunto de actividades, roles y documentos que deben ser contemplados para la ejecución del mismo. La descripción de los roles y documentos se presentan en las tablas 3 y 4 respectivamente, que serán utilizados para el análisis realizado en este trabajo.

Rol	**Competencia**
Cliente	Conocimiento de los procesos del cliente y capacidad de explicar los requisitos del cliente. Tiene la facultad de aprobar los requisitos y sus cambios. Conocimiento y experiencia en el dominio de aplicación
Líder de Proyecto	Capacidad de liderazgo. Experiencia en planificación, gestión de personal, delegación y supervisión, finanzas y desarrollo de software.
Equipo	Conocimiento y experiencia de acuerdo a su función en el proyecto. Conocimiento de las normas utilizadas por el cliente y/o por la PyMEs.
Analista	Conocimiento y experiencia de elicitar, especificar y analizar los requisitos. Conocimiento en diseño de interfaces de usuario y criterios ergonómicos. Conocimiento de las técnicas de revisión. Experiencia en el desarrollo y mantenimiento de software.
Desarrollador	Conocimiento y experiencia en los componentes de software y diseño de la arquitectura. Conocimiento de las técnicas de revisión. Conocimiento y experiencia en la planificación y realización de pruebas de integración.

	Experiencia en el desarrollo y mantenimiento de software.
Programador	Conocimiento y / o experiencia en la programación, integración y pruebas unitarias Conocimiento de las técnicas de revisión. Experiencia en el desarrollo y mantenimiento de software

Table 3. Roles ISO 29110

Documentos	Descripción
Declaración de trabajo	**Descripción del producto,** contiene: Propósito. Requerimientos generales. Alcance. Objetivos. Entregables.
Configuración del Software	**Identificación de conjunto de productos de software que se deben mantener actualizados,** contiene: Especificación de requerimientos, Diseño de software, Registro de trazabilidad, Software, Componentes, Casos de prueba, Reportes de pruebas, Manual de usuario, Documentación de mantenimiento.
Solicitud de cambio	**Documentación que identifica las solicitudes de cambios,** contiene: Propósito, estado de la solicitud, solicitante, impacto.
Plan de Proyecto	**Descripción de cómo el proyecto y sus actividades serán ejecutadas,** contiene: Descripción del producto, Propósito, Requerimientos generales, Alcance, Objetivos, Entregables, Tareas, Estimación de tiempo/costo/duración, Composición del equipo de trabajo, Riesgos.
Registro de aceptación	**Documento que establece la aceptación de los entregables por el cliente,** contiene: Registro de recepción de entregable, Fecha de recepción, Criterios de aceptación.
Minutas de reunión	**Registro de acuerdos establecidos con el cliente y/o equipo de trabajo,** contiene: Propósito de la reunión, asistentes, fecha, logros, cuestiones planteadas.

Table 4. Documentos ISO 29110

4 Roles, documentos y actividades de Scrum e ISO/IEC 29110

4.1 Roles

Analizando las competencias de los roles entre ISO/IEC 29110 y Scrum, se observa una relación directa para *Cliente*, *Líder de Proyecto* y *Equipo*, pero el resto de los roles del estándar están contemplados en las competencias del Scrum Team. En Scrum no se diferencian ni especifican las funciones de los integrantes del equipo, por lo cual los roles de *Analista*, *Desarrollador* y *Programador* no se pueden relacionar de forma directa. Ver tabla 5.

ISO 29110\ Scrum	Product Owner	Scrum Master	Scrum Team
Cliente	X		
Lider del Proyecto		X	

Equipo	X
Analista	*
Desarrollador	*
Programador	*

Table 5. - Comparación de roles

4.2 Documentos

Análogamente al análisis de roles, se estableció una comparación entre los documentos que solicita el estándar y los que utiliza la metodología. El estándar es muy riguroso en la definición de los documentos mientras que Scrum los maneja informalmente, por lo que, en algunos casos se podría establecer una mayor relación, dependiendo de la manera que sea construido.

En la tabla 6 se presentan las diferencias entre los documentos recomendados por el estándar y los que determina Scrum

ISO/IEC 29110	Scrum	Observaciones
Declaración de trabajo	Product Backlog	No posee una estructura definida, por lo tanto se puede acercar tanto como se desee al producto en cuestión.
Configuración del Software		Cada uno de los elementos de la configuración del software representan un producto en la norma, que no se corresponde con ningún artefacto en la metodología Scrum.
Solicitud de cambio	Sprint Backlog	Dado que en cada sprint puede incorporar modificaciones/mejoras en los requerimientos, es posible considerarlo una solicitud de cambio.
Plan de Proyecto	Product Backlog	No posee una estructura definida, por lo tanto se puede acercar tanto como se desee al producto en cuestión.
Registro de aceptación		Scrum no presenta un documento formal para registrar la aceptación de productos, pero en la práctica se deja constancia informal.
Minutas de reunión		Scrum no presenta un documento formal para registrar las minutas, pero se deja constancia de las reuniones Daily Scrum.

Table 6. - Comparación de documentos

4.3 Actividades

Como se mencionó anteriormente, ISO/IEC 29110 presenta dos grandes procesos (AP e IS) que abarcan todas las actividades a realizar durante el ciclo de vida de un proyecto.

En el proceso AP, las actividades de la etapa **Planificación de Proyecto**, equivalen a la recepción del **Product Owner** con la lista de requerimientos que se utiliza para

crear el **Product Backlog**. Las actividades de las etapas **Ejecución del plan de proyecto** y **Evaluación y Control del Proyecto** se relacionan con la ejecución del Sprint y las de la etapa **Cierre del Proyecto**, son equivalentes a la entrega final del proyecto. El Proceso de IS define actividades que están directamente ligadas al Sprint.

Actividades AP
Planificación del Proyecto
- *Revisar la Declaración de trabajo* - *Establecer tareas a realizar con dependencia y duración* - Establecer puntos de V&V - Definir equipo de trabajo con roles y responsabilidades - Definir capacitaciones - Estimar esfuerzo, costo y calendario -Identificar Riesgos
Ejecución del Plan de Proyecto
- Registrar el progreso del proyecto - *Analizar y evaluar los cambios y su impacto.* - Aprobar los cambios en el Plan. - *Mantener reuniones con el equipo de trabajo y el cliente.* - Actualizar el Repositorio
Evaluación y Control del Proyecto
- *Evaluar el progreso del Plan* *-Identificar y evaluar desviaciones y problemas de costo, calendario, técnicos.* -Documentar cambios y acciones correctivas. -Actualizar el Repositorio
Cierre del Proyecto
- *Realizar la entrega del producto según lo acordado.* - Realizar soporte al cliente - Finalizar el proyecto y firmar aceptación.

Actividad IS
Iniciación de la Implementación
- *Revisar el Plan de Proyecto con el equipo de trabajo y establecer tareas a realizar* - *Establecer el compromiso del equipo y el Lider* - Establecer el ambiente de Implementación
Análisis de Requerimientos de Soft
- *Revisar tareas asignadas* *-Elicitar, analizar y especificar requerimientos* - V&V los requerimientos - Control de versiones
Arquitectura y Diseño Detallado del software
- *Diseñar arquitectura. Componentes* - Rever especificación de requerimientos - Verificar Diseño y casos de prueba - Control de versiones
Construcción
- Rever el diseño para determinar secuencia de construcción. - *Codificar.* - Trazabilidad.
Prueba e integración
- Integrar componentes - *Realizar pruebas* y documentar - Verificar líneas base
Entrega
- Controlar Documentación - *Entrega del producto*

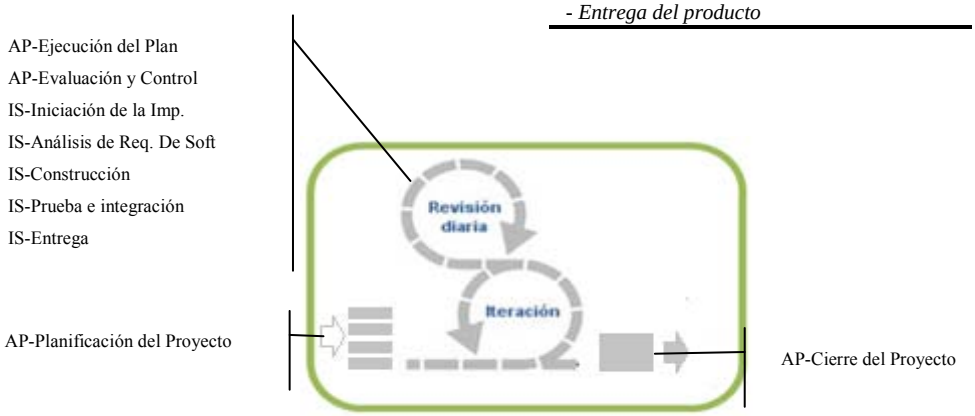


Fig. 2. Relación entre las actividades de Scrum y el estandar

En la figura 2 se observan las actividades de las etapas de cada proceso del estándar, y en cursiva se representan las actividades que son realizadas en Scrum.

5 Modelo Q-Scrum

Se presenta Q-Scrum, un modelo de procesos basado en Scrum, que integra un conjunto de roles, documentos, y actividades, capaz de generar los documentos necesarios para satisfacer los requisitos del estándar ISO/IEC 29110.

5.1 Roles

En la tabla 5 de la sección 4.1 se aprecia que con los roles de Scrum no es posible satisfacer los requisitos del estándar. Por lo que será necesario redefinir la estructura de roles. Q-Scrum propone crear un nuevo rol, Q-Scrum Analyst, que contemple las competencias de análisis y documentación que claramente son realizadas especifica y separadamente de las otras competencias, manteniendo en el Q-Scrum Team, las competencias de implementación (tanto de desarrolladores como programadores).

	Rol	Competencias
QPO	Q-Product Owner	Product Owner / Cliente
QSM	Q-Scrum Master	Scrum Master / Líder de Proyecto
QST	Q-Scrum Team	Scrum Team / Equipo- Programadores – Desarrolladores
QSA	Q-Scrum Analyst	Analista

Table 7. Roles Q-Scrum

5.2 Documentos

En 4.2 se presentaron las diferencias entre los documentos de Scrum y el estándar, donde es evidente que la documentación generada por Scrum es insuficiente para satisfacer el estándar. Q-Scrum propone estructurar los documentos Product Backlog y Sprint Backlog, sin perder la flexibilidad de Scrum, formalizar los documentos de Registro de Aceptación y Minutas de reunión, que son utilizados habitualmente en desarrollos Scrum, e incorporar el documento de Configuración del software.

	Q-Scrum	**Observaciones**
dQPB	Q-Product Backlog	Plantilla básica, con la información requerida por la descripción de trabajo y el plan de proyecto del estándar, que inicialmente se completó con la información básica para iniciar el proyecto y se fue actualizando a lo largo del desarrollo.
dQSC	Q-Software Configuration	Documento donde se irán incorporando todos los registros de las actividades realizadas.

dQSB	Q-Sprint Backlog	Plantilla básica, que incluye las solicitudes de cambio, donde se incorporarán los requerimientos de cada sprint.
dQAR	Q-Acepted Record	Plantilla básica donde se registrará la aceptación de los productos
dQMR	Q-Meeting Record	Plantilla básica donde se registrarán las decisiones de las reuniones como por ejemplo las reuniones Daily Scrum.

Table 8. Documentos Q-Scrum

5.3 Actividades

En base a las relaciones entre las actividades descritas en 4.3, Q-Scrum propone modificar los procesos de AP y IS para soportar la nueva estructura de roles y generar/mantener los documentos de Q-Scrum.
El proceso de AP quedaría compuesto por las etapas **Inicio**, **Planificación de Proyecto**, **Ejecución y Evaluación de proyecto** y **Cierre**. El Proceso de IS con las etapas **Iniciación de la Implementación y Análisis de requerimientos preliminar**, **Ejecución** y **Pre-Entrega**. La Figura 3 presenta la estructura y relación de los procesos del modelo propuesto.

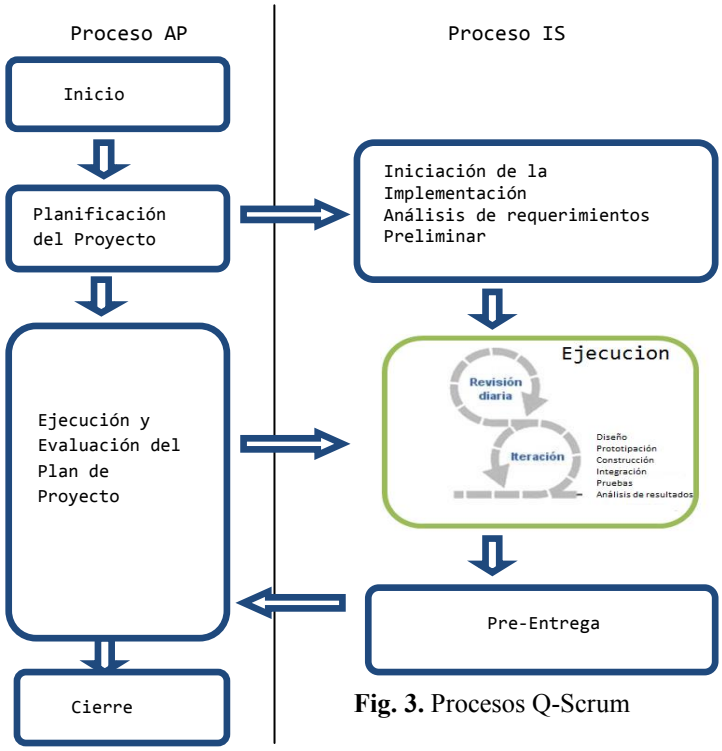


Fig. 3. Procesos Q-Scrum

Procesos AP.

Inicio.

Se recibe una petición del QPO, el QSM da inicio al proyecto generando el dQPB, se asigna un QSA.

Planificación del Proyecto.

El QSM y el QSA tomarán el dQPM preliminar y analizarán la factibilidad, riesgos, tareas a realizar, dependencias, duración, puntos de VyV, estimarán esfuerzos, costos y calendario. En base a eso asignarán un QST. El QSM actualizará el sQPM y el QSA generará el dQCS.
El QST ya iniciará el proceso de IS.

Ejecución y Evaluación del Plan de Proyecto.

El QSM y QSA registrarán y evaluarán el progreso del proyecto, analizarán posibles cambios, correcciones de costos y calendarios, ante cada spring.
El QSA registrará los cambios en el dQCS.

Cierre.

El QSM registrará la entrega final con un dQAR y finalizará el proyecto.

Proceso de IS.

Iniciación de la Implementación y Análisis de requerimientos preliminar.

El QSM y el QST realizarán la primera reunión generando el dQSB del primer sprint. El QSA formalizará la reunión en un dQMR.

Ejecución.

Cada iteración del Sprint realizará las siguientes sub-etapas:

Actividad	Rol	Registro
Análisis		
- Revisar tareas asignadas	QSM	dQPB y dQMR
- Elicitar, analizar y especificar requerimientos	QSA	dQSB
- V&V de los requerimientos	QSM	dQPB y dQCS
Diseño y Prototipado		
- Diseñar arquitectura. Componentes	QST	dQSB y dQPB
- Rever especificación de req	QSM	dQSB y dQPB
- Verificar Diseño y casos de prueba	QSM	dQPB y dQCS
Construcción		
- Codificar.	QST	
- Trazabilidad	QSA	dQCS
Prueba e integración		

-Integrar componentes	QST	dQSB
- Realizar pruebas y documentar	QST	dQSB
- Verificar líneas base	QSA	dQCS
Pre – Entrega		
-Controlar Documentación	QST	dQCS
-Pre - Entrega del producto	QST	dQCS

Entrega.

Cerrada la iteración, el QSM registrará la entrega en un dQAR y el QSA actualizará el dQCS.

6 Conclusiones

Se presentó una descripción de Scrum y del estándar ISO/IEC 29110 y una relación entre ellos en función de los roles, documentos y actividades, concluyendo que la metodología Scrum aplicada de forma directa no es capaz de satisfacer los requisitos del estándar.

Se propone Q-Scrum, agregando a Scrum el rol analista (QSA), incorporando un documento para la gestión de la configuración del software (dQSC) y formalizando las minutas (dQMR) y los documentos de aceptación (dQAR). Se fusionaron las actividades de Scrum con los procesos de AP e IS, generando un modelo de procesos capaz de satisfacer los requisitos del estándar.

Se ha iniciado la aplicación del modelo Q-Scrum en PyMEs de la región, que desarrollan bajo la metodología Scrum y poseen intenciones de lograr una mejora de proceso. La retroalimentación recibida de estas experiencias permitirá ajustar el modelo propuesto.

7 Referencias

[1] Comunidad Latinoamericana de Metodologías Ágiles http://www.agiles.org Julio 2013

[2] Muñoz, Oktaba, "Especialización de MoProSoft basada en el método ágil Scrum", Editorial académica Española Año 2011

[3] A. Pasini, S. Esponda, P. Pesado and R. Bertone., Aseguramiento de calidad en PYMES que desarrollan software. una experiencia desde el proyecto COMPETISOFT. 2008. pp. 957-966.

[4] Piattini, Oktaba, Orozco, "COMPETISOFT. Mejora de procesos software para pequeñas y medianas empresas", Editorial Ra-Ma, Año 2008

[5] ISO/IEC 29110:2011, "Software engineering -- Lifecycle profiles for Very Small Entities (VSEs)" 2011, ISO

[6] Henrik Kniberg “Scrum y xp desde las trincheras” (2007) Libro Online InfoQ http://www.infoq.com/minibooks/scrum-xp-from-the-trenches, julio 2013

[7] Schwaber, Ken Agile Project Management with scrum, Redmon Wshington: Microsoft Press 2004

[8]Pablo Lledó , Gestión Ágil de Proyectos – Pablo Lledó. ISBN: 978-1-4669-2119-1, Traford Published, 2012.

Inserción del mantenimiento en los procesos ágiles

Karla Mendes Calo[1] Karina Cenci[1] Pablo Fillottrani[1,2]

[1]Laboratorio de Investigación en Ingeniería de Software y Sistemas de Información
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
[2]Comisión de Investigaciones Científicas, Provincia de Buenos Aires
{kmca,kmc,prf}@cs.uns.edu.ar

Resumen El mantenimiento del software abarca todas las actividades asociadas con el proceso de cambio del software. Los artefactos generados durante la etapa de desarrollo, son utilizados como soporte en el mantenimiento de las aplicaciones. Las metodologías ágiles valoran y promueven la comunicación verbal entre los integrantes del equipo por sobre la documentación. Frecuentemente esto origina errores o demoras involuntarias en el mantenimiento, en especial, si decisiones relevantes no están registradas.
En este trabajo, se expone una práctica no ágil, que se adapta al modelo ágil en la etapa de mantenimiento, la cual consiste en una Iteración de Reordenamiento del Conocimiento (IRC) que se lleva a cabo antes de iniciar el proceso de mantenimiento ágil, recomendando la utilización de esta práctica en proyectos que superan el año de desarrollo, y se utilizó una metodología ágil para el proceso. Se muestra cómo la combinación y adaptación de ciertas prácticas utilizadas en metodologías tradicionales con metodologías ágiles, permite obtener procesos híbridos, con los beneficios de ambas metodologías.

Palabras claves: Mantenimiento - Procesos Ágiles - Gestión Conocimiento

1. Introducción

Tradicionalmente, el proceso de desarrollo de software ha sido descripto como un proceso secuencial. Un modelo secuencial extensamente conocido es el modelo cascada, en la que una fase puede ser iniciada solo si todas las fases precedentes fueron completadas. Como contraste de este proceso, las metodologías ágiles se caracterizan por realizar entregas de software al finalizar cada iteración, cuya duración típicamente no excede el mes. En cada iteración se incluyen todas las fases del desarrollo de software.

El término desarrollo de software ágil fue introducido en 2001 por el Manifiesto Agil [3], y desde entonces, se han publicado varios artículos sobre el tema. Esta metodología enfatiza la simplicidad y la comunicación, tanto entre los miembros del equipo y con el cliente, minimizando la documentación formal. Cambios en los requerimientos durante las iteraciones no solo son esperados, sino que son fomentados, posibilitando cambiar de

dirección rápidamente de acuerdo a las necesidades del cliente, impulsando el concepto de agilidad.

La información, el conocimiento y la gestión del conocimiento son fundamentales para el desarrollo de cualquier proyecto de software. Para el caso de metodologías ágiles, la información se puede encontrar en distintos tipos de documentos: e-mails entre el cliente y miembros del equipo, posteos acerca de discusiones técnicas con alternativas para su resolución, lecciones aprendidas, historias gestionadas por un grupo de procesos centrales. Dado que el proceso de construcción de software utilizando metodologías ágiles es un proceso cíclico e iterativo, hay que tener en cuenta que la gestión de este conocimiento, implica asegurar la actualización y distribución del conocimiento entre los miembros del equipo, y combinarlo además con nuevo conocimiento que se irá sumando a través de las iteraciones durante el proceso de desarrollo.

Si bien la entrega del producto completo es el hito más importante al que un equipo de desarrollo aspira, las tareas en el proceso de desarrollo de software no terminan allí. Inevitablemente sufren cambios para su permanencia y utilidad. Aparecen nuevos requerimientos por parte de los usuarios que involucrará el desarrollo de nuevas funcionalidades, o bien por cambios en normativas legales e impositivas, cambios tecnológicos, errores detectados en su funcionamiento, mejoras en la performance, mejoras en el rendimiento. Estos cambios en algunos casos pueden involucrar reingeniería de alguna parte del producto. A esta etapa del desarrollo de software, lo denominamos mantenimiento.

El resto del trabajo está organizado de la siguiente manera. En la sección 2, se introduce conceptos, políticas y tipos de mantenimiento. Sección 3, son presentadas las características del mantenimiento en las metodologías ágiles. Sección 4 presenta un caso de estudio y en sección 5 experiencias de mantenimiento en las metodologías ágiles. La propuesta de una metodología híbrida para el mantenimiento, se presenta en la sección 6, y por último conclusiones.

2. Mantenimiento

El mantenimiento del software no es como el mantenimiento del hardware, que es la devolución del artículo (ítem) a su estado original. El mantenimiento del software desplaza un artículo de su estado original. El mismo abarca todas las actividades asociadas con el proceso de cambio del software. Esto incluye todo lo asociado con correcciones de error (*bug*), mejoras funcionales y de rendimiento, proporcionar compatibilidad con versiones anteriores, actualización del algoritmo, la creación de métodos de acceso a la interfaz de usuario, y cualquier otro cambio.

Mantenimiento de Software es definido en el standard del IEEE para el Mantenimiento de Software, IEEE 1219, como la modificación de un producto de software después de entregado para corregir fallas/defectos, para mejorar el rendimiento u otros atributos, o para adaptar el producto a los cambios del ambiente.

En software, agregar una autopista de seis carriles de autos de un puente del ferrocarril es considerado mantenimiento, y es particularmente valioso si el mismo se puede realizar sin frenar el tráfico de los trenes. El desafío es el diseño de software de tal manera que el mantenimiento se pueda realizar sin frenar al software que está en producción.

Las políticas sobre el mantenimiento presentadas en [9] son las siguientes: a) *Tradicional* No considera la posibilidad de mantenimiento. b) *Nunca* Decide que nunca va a ocurrir mantenimiento. Simplemente se escriben muy buenos programas correctos

desde el inicio. c) *Discreto* En teoría, el proceso acepta el hecho del cambio, se mantienen listas de partes y herramientas sobre cada ítem, los cambios son realizados bajos estrictos controles. d) *Continuo* El cambio es constante. La migración de hardware, software y comportamiento durante el funcionamiento del sistema es necesaria.

En los desarrollos actuales de productos de software, se acepta el hecho del cambio, pero no siempre el diseño y arquitectura están preparados para soportar las diversas modificaciones sin comprometer las calidades (rendimiento, confiabilidad, etc) del mismo. En función de la clase de cambio que deba llevarse a cabo, se puede clasificar el mantenimiento como a) *Adaptativo* Modificaciones para adaptar el producto del software a los cambios en los requerimientos de datos y procesamiento del ambiente (entorno). [2], b) *Preventivo* Modificaciones al producto de software después de entregado para detectar y corregir fallas latentes antes de que se conviertan en fallas funcionales. [4], c) *Correctivo* Tiene como objetivo solventar una deficiencia en un componente del sistema de información (puede ser software o documental). Entiéndase deficiencia como algo que debería funcionar o estar correcto y que no lo está., d) *Perfectivo* Modificaciones al producto de software después de entregado para detectar y corregir fallas latentes en el producto de software antes de que se manifiesten como fallas. [4], e) *Ayuda al Usuario* Responde a las demandas de los usuarios distintas de las adaptativas, correctivas, preventivas o perfectivas. [2].

Existen algunos problemas que son específicos del mantenimiento, como es la necesidad de descubrir decisiones de diseño de alto nivel, alto volumen de información a considerar, necesidad de entrenamiento de nuevo personal, resistencia al cambio por desgaste de la arquitectura.

En las metodologías tradicionales, un aspecto significativo que gobierna el proceso es el exhaustivo uso de artefactos, para registrar el proceso y la documentación. La documentación, si es correcta, completa y consistente, ha sido considerada como una poderosa herramienta para los ingenieros de software para alcanzar el éxito. En contrapartida, una pobre documentación es considerada la principal razón para la rápida degradación de la calidad del software y el envejecimiento. El propósito de la documentación no es solamente describir el sistema de software sino también registrar el proceso. Los ingenieros de software necesitan poseer un buen conocimiento del sistema para estar habilitados a continuar con la evolución del mismo.

En las metodologías tradicionales, el mantenimiento se apoya en gran medida en los artefactos generados en etapas de análisis y diseño. Estos artefactos generados son el soporte fundamental del conocimiento para llevar a cabo el mantenimiento de las aplicaciones.

La figura 1 muestra cómo es el despliegue de un producto, considerando en el desarrollo todas las etapas necesarias para construir el producto de software, a partir de la entrega del mismo se pasa a la etapa de mantenimiento que se encarga de todos los tipos de modificaciones requeridas.

3. Mantenimiento de Software en las Metodologías Ágiles

La documentación en ambientes ágiles no está registrada formalmente, la transferencia es informal y primordialmente en forma oral. Se considera a la documentación como un aspecto secundario, centrando el proceso en el producto.

En ambientes ágiles, un riesgo significativo es que la mayoría de la documentación está en la memoria de los desarrolladores o en posters temporarios. Con la ausencia de una apropiada documentación, existe un alto riesgo que el conocimiento organizacional

Figura 1. Ciclo de Despliegue

de un sistema sea olvidado, mal interpretado o perdido. Una alta rotación del personal puede conducir a una pérdida significativa del conocimiento del sistema si un miembro habilidoso/dotado abandona y su conocimiento y experiencia no está registrada. Además, existen dificultades al momento de entrenar a nuevos contratados cuando la documentación se encuentra en el código.[5]

La figura 2 muestra el despliegue de un producto utilizando una metodología ágil, el desarrollo del mismo está gobernado por las iteraciones, donde cada una de las iteraciones genera un hito, que es un subconjunto de requerimientos desarrollados para poner a producción. A partir del momento que el producto está en producción pueden aparecer defectos en el mismo que hay que corregir, el modelo de proceso está gobernado por los cambios permanentes a través de nuevos requerimientos y de adaptaciones al producto en funcionamiento.

Figura 2. Ciclo de Despliegue

Las metodologías ágiles son adoptadas cada vez con más frecuencia no solo como ciclo de vida en la etapa de desarrollo, sino que también tienden a utilizarse en el proceso de mantenimiento, ya que el uso de distintos procesos en una organización generalmente disminuye la eficiencia del negocio [7] [10]. En función de esto, hay una necesidad de revisar el mantenimiento en un contexto ágil.

Svensson y Host [11] presentaron un caso de estudio de un mantenimiento de software, y los resultados mostraron que un gran número de prácticas ágiles tales como *planning game*, programación por pares y la integración continua son apropiados en el contexto del mantenimiento y facilitan la transición del conocimiento. Otros trabajos sobre el mismo aspecto son Shaw [8] y Rico [6].

La adopción de metodologías ágiles en la etapa de mantenimiento debe ser llevado a cabo con cierta cautela, debido justamente a los escasos artefactos generados durante las etapas de desarrollo, que sirvan de soporte en el mantenimiento de las aplicaciones.

Consideraremos que el alcance de las responsabilidades del equipo ágil en el proceso como tareas de mantenimiento en cada iteración, puede incluir la corrección de defectos, llevar a cabo mejoras y resolver problemas de soporte en general.

4. Caso de Estudio

En este trabajo vamos a considerar el desarrollo de software en el que participan equipos de entre 10 y 12 personas durante un período de entre 10 meses y 24 meses. La aplicación tiene una complejidad de desarrollo media-alta, esto debido tanto al origen del negocio como a la tecnología utilizada. Durante el período de desarrollo se utilizó una metodología de desarrollo ágil.

Si bien una de las características de las metodologías ágiles es la falta de énfasis en la documentación formal y mayor énfasis en la comunicación entre los miembros del equipo y el cliente, debido al tamaño del proyecto y al número de integrantes del equipo, se integró una herramienta colaborativa de gestión de conocimiento, donde se definieron actividades, artefactos, documentación y notaciones a ser utilizadas, orden de ejecución de las actividades que ayudaron desde el inicio hasta la finalización en la construcción del producto.

Como todo proceso evolutivo, a medida que se avanzaba en el desarrollo, se realizaron cambios sobre funcionalidades ya implementadas, haciendo que en algunos casos la funcionalidad se comporte de diferente manera y en algunos casos, entregando resultados diferentes. Todos los cambios, defectos encontrados y mejoras, fueron siendo documentados en la herramienta de gestión de conocimiento.

Una vez finalizada la etapa de desarrollo y con el producto finalizado y en funcionamiento, se decidió llevar a cabo un período de mantenimiento utilizando también un proceso ágil. Para ello, tuvieron que ajustarse algunas pautas o reglas al momento de realizar mantenimiento evolutivo de software, propias de las metodologías ágiles. Uno de los objetivos consistía en colaborar con el equipo de trabajo para que puedan tomar las decisiones correctas, con la mayor cantidad de información posible.

Experiencia realizada con el mantenimiento Una lista de factores que se tuvieron en cuenta en el proceso de mantenimiento del caso de estudio propuesto, sobre los que hubo que realizar algunos ajustes o adaptaciones, sin comprometer el espíritu del modelo de proceso ágil respetando las características de incremental, cooperativo, sencillo y adaptativo tal como lo definió Abrahamsson [1]. Los aspectos considerados son los que se detallan a continuación:

1. Rotación de recursos en el equipo: Durante el ciclo de vida del sistema de software, el conocimiento obtenido con la experiencia de los desarrolladores y quienes han tenido la responsabilidad de mantenerlos en funcionamiento, se pierde una vez que estos dejan la organización o son asignados a otros proyectos. Si consideramos que con frecuencia es personal sin experiencia el asignado a las nuevas tareas de Mantenimiento de Software, es un factor a considerar al momento de armar el equipo ágil de mantenimiento.
2. Falta de conocimiento del negocio, diseño y arquitectura. Falta de conocimiento explícito y peor aún tácito, por parte del equipo de mantenimiento. Esto puede deberse a la falta de experiencia de los integrantes del equipo, pobre traspaso de

información por parte de integrantes que ya no pertenecen, sumada a una escasa documentación, o distribuida en diferentes documentos o diferentes herramientas de gestión de conocimiento.
3. Escasa documentación: La mayoría de la información del producto y sus características la conoce el equipo de desarrollo, pero no está registrada en algún documento de diseño.
4. Reuniones periódicas o por demanda: Estas reuniones reemplazan a las reuniones diarias, ya que se trabaja con un equipo reducido con respecto al original y también porque parte de las tareas a realizar en la iteración actual son de corta duración.
5. Cambios en priorización en la iteración actual: (Mantenimiento Correctivo vs Evolutivo): Cuando existe un solo equipo de trabajo, si se detecta algún defecto en la aplicación que está en producción, (teniendo en cuenta la criticidad y el impacto del mismo), alguno de los integrantes del equipo debe dejar pendiente las tareas relacionadas con la iteración actual, y llevar a cabo la corrección, y luego de las pruebas correspondientes, liberar la nueva versión al ambiente de producción. Estas cuestiones pueden demorar la iteración actual o bien llegar pero no con la calidad que se pretende. Se valora en la etapa de mantenimiento, la rápida y eficiente reacción ante cambios abruptos de prioridades como consecuencia de un defecto de alto impacto y criticidad en el producto final.
6. Duración de las iteraciones: La duración, puede variar de iteración a iteración. Puede ocurrir que, relacionado con el punto anterior, haya un cambio de prioridad ante un defecto detectado en el producto en producción, y deba liberarse una nueva versión del producto (*release*) con la solución de ese defecto en cualquier punto de la iteración actual.
7. Iteraciones vs Releases: Coordinación entre los integrantes del equipo en los pasajes a producción, ocasionados por el mantenimiento correctivo en cualquier momento de la iteración; marcado esto por la criticidad del defecto (*release*), con aquellos que corresponden con las tareas planificadas para la iteración actual, como parte de la evolución del sistema. Deben tenerse en cuenta los *branches* en los que se trabaja, ser cuidadosos con las fusiones del código.
8. Antes de cada pasaje a producción del producto, debe llevarse a cabo una regresión, y verificar que no se han introducido nuevos defectos, asegurando la calidad del producto.

Si bien uno de los principios del Manifiesto Ágil [3] expresa "Desarrollar software que funciona más que conseguir una buena documentación", es necesario comprender que se debe diferenciar la construcción de modelos con el único fin de documentar el producto generado, de la elaboración de modelos como proceso de diseño de soluciones.

En el caso particular del caso estudiado, debido al tamaño y complejidad del mismo en la etapa de desarrollo, se generó documentación, soportadas en una herramienta de gestión de conocimiento colaborativa. Se propuso no documentar la aplicación completa, sino solamente cuando fueran útiles y necesarias para definir la solución adecuada. Es decir, lineamientos y actividades que faciliten la construcción del producto a partir de la creación de historias de usuario funcionales y no funcionales, wikies, tareas, defectos encontrados, sin perder las características enunciadas en el Manifiesto Ágil [3] y como manera de dar soporte posteriormente a las actividades de mantenimiento.

El desafío más grande es ser lo suficientemente maduros como organización para asumir como propias las prácticas ágiles y evolucionar la forma de trabajo adaptándose a las exigencias cambiantes del entorno sin descuidar la calidad de la entrega en cada iteración.

Un problema significativo fue la recuperación de la información y todo el *background* generado durante el desarrollo del producto. Este es un punto crítico, ya que podemos encontrarnos con diferentes obstáculos que impidan, retrasen o haga en el peor de los casos, que se agreguen involuntariamente defectos en la aplicación que ya está en producción. En especial teniendo en cuenta que no todo el equipo original que llevó a cabo el desarrollo, es el asignado en la etapa de mantenimiento.

5. IRC y Mantenimiento

La figura 3 muestra las actividades involucradas en el mantenimiento ágil de software para proyectos que tienen un período de duración superior al año, utilizado en el caso de estudio en cuestión. Se explicará aquí qué decisiones se tomaron para asegurar el menor impacto posible sobre el producto ya liberado a producción durante la etapa de mantenimiento.

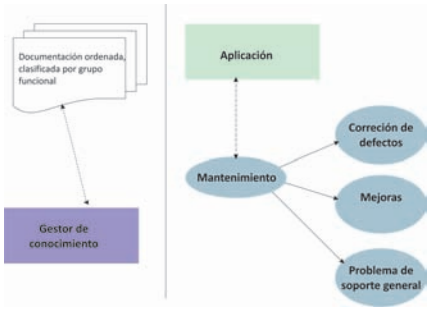


Figura 3. Mantenimiento en el desarrollo ágil

El lado derecho de la figura 3 muestra cuáles son las responsabilidades del equipo de mantenimiento que fueron divididas en tres categorías:

- Corrección Defectos: son aquellos errores que pueden ser causados por fallas en el código, por datos erróneos, o por cualquier número de factores externos.
- Mejoras: Son requerimientos solicitados por los usuarios finales, que surgen del uso del producto, y por ello aparecen nuevas necesidades, estas pueden ser modificaciones del producto ó nuevas funcionalidades, que les permitirán obtener mejores resultados en el uso diario del producto generado.
- Problemas de soporte general: Esta categoría abarca todo aquello que no es cubierto por las dos categorías anteriores. Por ejemplo análisis de performance, consultas a la base de datos, manejo de la entrega (*delivery*) de nuevas versiones al cliente, entre otros.

El lado izquierdo de la figura 3 muestra la incorporación de la herramienta utilizada en la gestión de conocimiento, en la que el equipo ágil soporta las tareas de mantenimiento. En esta herramienta se encuentran las historias de usuario, wikies, decisiones

del diseño, diseños de casos de test, documentación con los pasos para realizar los pasajes a producción, y todas las tareas que formaron parte de los pedidos pendientes (*backlog*) del desarrollo del producto.

Debido a la velocidad con la que se trabaja utilizando estas metodologías, y teniendo en cuenta que entre los valores del Manifiesto Agil [3] se promueve la comunicación verbal entre los integrantes del equipo por sobre la documentación, y que hay que dar respuesta a los cambios, sobre el cumplimiento de un plan, puede esto convertirse en una desventaja, ya que la información del producto y sus características existen en la mente de los desarrolladores, pero no quedan en un documento de diseño, con algún tipo de ordenamiento que pueda ser utilizado posteriormente por un equipo de mantenimiento.

Este tipo de situaciones, puede generar en los integrantes del equipo cierta incertidumbre, falta de seguridad y confianza al momento de realizar por ejemplo, un cambio en el diseño, efectuar una mejora, solucionar un defecto o estimar tareas que pueden poner en riesgo el éxito en esta etapa. Decisiones tomadas durante el diseño acompañado por su escasa de documentación, pueden llegar a confundir o desconcertar al equipo de mantenimiento, al no comprender el porqué de ciertas decisiones que fueron tomadas. Esto suele ocurrir con equipos de mantenimiento en los que la mayoría de sus integrantes no formaron parte del equipo de desarrollo original.

Para mitigar esta situación, se propuso a manera de experiencia, organizar, ordenar y reagrupar el conocimiento repartido entre el gestor de conocimiento, mails, documentación en general del proyecto.

La propuesta se basa en que, antes de iniciar el proceso de mantenimiento propiamente dicho, se lleve a cabo una iteración denominada Iteración de Reordenamiento del Conocimiento (IRC). La Figura 4 en el lado izquierdo muestra los incrementos realizados durante el período de desarrollo del sistema, los despliegues realizados en cada iteración, donde en cada incremento se tienen en cuenta los requerimientos iniciales para desarrollar el producto, y los nuevos requerimientos que van surgiendo a medida que se evoluciona para llegar al producto final. La Figura 4 muestra cómo se integra esta práctica no ágil (IRC), al proceso de mantenimiento ágil, como una iteración de transición entre la finalización de la etapa de desarrollo y el inicio de la etapa de mantenimiento.

La duración de esta iteración debe tomar entre una y dos semanas aproximadamente. Su objetivo es el de organizar, depurar la información que se ha registrado en el gestor de conocimiento, incorporar aquella información que no se encontraba accesible, completar información relevante de algunas wikies, marcar las historias de usuario como obsoletas, eliminar información confusa, vincular tareas que han tenido relación entre sí, agregar etiquetas a las wikis de manera de organizarlas por grupo funcional. Esto facilita futuras búsquedas, haciéndolas más efectivas en cuanto a los resultados que se obtienen.

El tiempo invertido en la reestructuración del conocimiento, fue capitalizado durante el período de mantenimiento propiamente dicho, ya que se logró que en cambios y mejoras de algunas funcionalidades, e incluso de diseño, se recuperara la información requerida, las wikies actualizadas con información confiable, como base para comprender el alcance, el impacto y el riesgo en los cambios y mejoras propuestas.

La IRC se incorpora como una práctica no ágil, implementando una adaptación híbrida en el proceso de mantenimiento ágil de software. Comprobamos que en un desarrollo ágil que llevó más de un año en su implementación, durante el período de mantenimiento se redujeron algunos efectos negativos. Decisiones de cambios en el diseño, cambios en funcionalidades, mejoras de performance, fueron tomadas sobre una

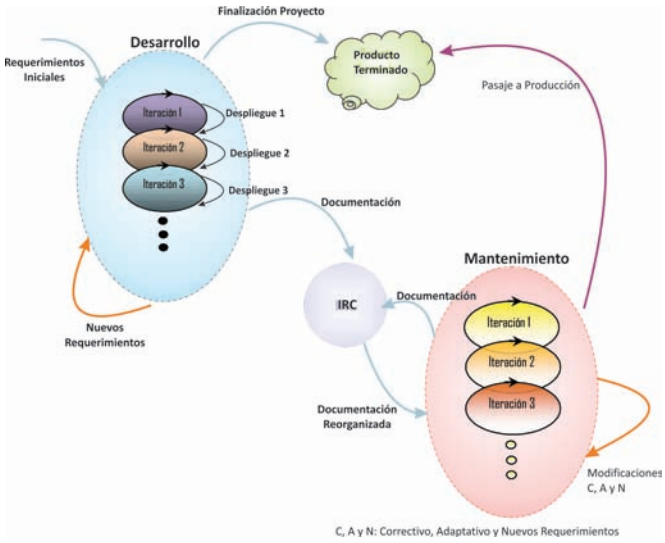


Figura 4. Mantenimiento con IRC

base más sólida de conocimiento documentado, reagrupado y depurado al implementar IRC. Se redujo el impacto y riesgo de algunos cambios durante el mantenimiento, sin perjuicio de perder la agilidad del proceso.

6. Conclusiones

La industria de desarrollo de software sufre retrasos en la finalización de proyectos, debido a los requerimientos de documentación pesados de los modelos de procesos tradicionales. Como contrapartida, los modelos de desarrollo ágil se caracterizan por el poco peso que tiene la documentación *detallada* durante el proceso de desarrollo, y la importancia de la intercomunicación entre los integrantes del equipo. Hemos descripto anteriormente cómo en el gestor de conocimiento, la información suele quedar desactualizada, insuficiente o incompleta debido a la velocidad con que se trabaja en estas metodologías y a los frecuentes cambios en los requerimientos. Esta situación termina impactando en el proceso de mantenimiento de software utilizando metodologías ágiles, sobre proyectos que tienen una duración de más de una año. El entendimiento de la traza del código, termina siendo uno de los principales puntos en la etapa de mantenimiento del software, y en ocasiones acompañado por la falta de comprensión del porqué de ciertas decisiones tomadas en su diseño.

En este trabajo, se presentó una práctica no ágil, que adaptamos al modelo ágil. Consiste en una Iteración de Reordenamiento del Conocimiento (IRC) que se lleva a cabo antes de iniciar el proceso de mantenimiento. La propuesta consiste en reorganizar el conocimiento clave, que aporte valor directo al producto, sin llegar a escribir ni reescribir requerimientos detallados. El uso de esta práctica híbrida antes de iniciar la etapa de mantenimiento, permitió que el equipo alcanzara un buen rendimiento, sobre todo en lo concerniente a cambios en las funcionalidades, cambios en el diseño, cambios en

la arquitectura, implementación de nuevas funcionalidades, donde resultó más sencillo encontrar información relacionada, que se encontraba en el gestor de conocimiento. La combinación y adaptación de ciertas prácticas utilizadas en metodologías tradicionales con metodologías ágiles, permiten obtener procesos híbridos, con los beneficios de ambas metodologías. Las organizaciones dedicadas al desarrollo de software, no siempre deben adoptar el modelo ágil en sus procesos, sino que deben adaptarlo a sus procesos y necesidades.

Si bien no se recomienda utilizar metodologías ágiles en desarrollos que excedan los 8-12 meses, la industria del software las adopta cada vez con mayor frecuencia, sin importar la duración de los proyectos. Esta práctica híbrida (IRC), podría adoptarse durante este proceso cada cierto período de tiempo (mayor a ocho meses), también por demanda del equipo. Esta decisión puede tomarse si se tienen errores en el diseño, o se introducen defectos que impactan en el comportamiento del producto, debido a la falta de conocimiento de ciertos detalles del diseño, la implementación o el negocio.

Referencias

1. P. Abrahamsson, O. Salo, J. Ronkainen, and J. Warsta. *Agile Software Development Methods - Review and Analysis.* VTT Elektroniikka, 2002.
2. A. Abran and H. Nguyenkim. Measurement of the maintenance process from demand-based perspective. *Journal of Software Maintenance. Research and Practice.*, 5(2):63–90, 1993.
3. K. Beck, J. Grenning, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, S. Mellor, K. Schwaber, J. Sutherland, and D. Thomas. *Agile Manifesto*, 2001.
4. International Organisation for Standarization. *International Organisation for Standarization. Software Engineering - Software Life Cycle Processes - Maintenance, ISO/IEC Standard 14764.*, 2006. International Organisation for Standarization: Geneva, Switzerland.
5. M. Kajko-Mattsson. Problems in agile trenches. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*, ESEM '08, pages 111–119. ACM, 2008.
6. D. Rico. *Agile Methods and Software Maintenance.*, 2008. http://davidfrico.com/rico08f.pdf.
7. P. M. Senge. *The Fifth Discipline: The Art & Practice of The Learning Organization.* Doubleday, 1990/2006.
8. S. Shaw. Using agile practices in a maintenance environment. *Intelliware Development Inc*, 2007.
9. P. Stachour and D. Collier-Brown. You don't know jack about software maintenance. *Communications of the ACM.*, 52(11):54–58, 2009.
10. H. Svensson and M. Host. Introducing an agile process in a software maintenance and evolution organization. In *Proceedings of the Ninth European Conference on Software Maintenance and Reengineering*, CSMR '05, pages 256–264. IEEE Computer Society, 2005.
11. H. Svensson and M. Host. Introducing an agile process in a software maintenance and evolution organization. In *Proceedings of the Ninth European Conference on Software Maintenance and Reengineering*, CSMR '05, pages 256–264. IEEE Computer Society, 2005.

Trazabilidad de Procesos Ágiles: un Modelo para la Trazabilidad de Procesos Scrum

Roberto Nazareno[1, 2], Horacio Leone[1, 3], Silvio Gonnet[1,3]

[1] INGAR (CONICET – UTN). Avellaneda 3657, Santa Fe, Argentina
[2] Universidad Nacional de La Rioja, La Rioja, Argentina
[3] Facultad Regional Santa Fe, Universidad Tecnológica Nacional, Santa Fe, Argentina
{rnazareno, hleone, sgonnet}@santafe-conicet.gov.ar

Abstract. La trazabilidad es considerada en metodologías ágiles como un aspecto esencial a incorporar para la producción de software de calidad. Sin embargo, los procesos de desarrollo ágiles en contraposición a los procesos de desarrollo "pesados", no permiten la aplicación directa de las técnicas de trazabilidad tradicionales. En consecuencia, es fundamental desarrollar modelos que permitan trazar los requerimientos bajo el enfoque de los métodos ágiles. En este trabajo se aborda esta problemática centrada en la metodología ágil Scrum. El modelo propuesto es desarrollado con el objetivo de brindar soporte a las siguientes preguntas de competencia: i) ¿Qué eventos originaron un artefacto en particular?; ii) ¿Qué requerimientos guiaron la generación de tal artefacto?; ¿Quiénes son los participantes involucrados en un evento dado? Las respuestas a estar preguntas asistirían las tareas de trazabilidad en proyectos ágiles tales como: Stakeholders con Requerimientos, User Stories con Versiones y Requerimientos con Requerimientos.

Keywords: Scrum, Trazabilidad, Procesos Ágiles

1 Introducción

En los últimos veinte años surgieron diversas propuestas para brindar soporte a la trazabilidad de requerimientos a lo largo del ciclo de vida del sistema. Se entiende por trazabilidad: (i) el grado en el cual se puede establecer una relación entre dos o más productos del proceso de desarrollo, especialmente productos que posean una relación de predecesor-sucesor o superior-subordinado; o (ii) el grado en el cual se puede establecer la razón de la existencia de cada elemento en un proceso de desarrollo de software [1]. Ambas definiciones de trazabilidad explicitan una relación entre artefactos del proceso de desarrollo y su adopción provee un soporte esencial a la producción de software de calidad. En estas dos últimas décadas, las metodologías de desarrollo tradicionales o "pesados" centraron sus prácticas de trazabilidad en el establecimiento de trazas desde requerimientos a otros artefactos de desarrollo. Siendo una traza una relación entre dos o más productos del proceso de desarrollo [1].

La trazabilidad es también considerada en metodologías ágiles como un aspecto fundamental a estudiar para desarrollar sistemas de calidad [2][3]. Sin embargo, los procesos de desarrollo ágiles difieren de los procesos de desarrollo pesados, no permitiendo la aplicación directa de las técnicas de trazabilidad tradicionales en los métodos ágiles.

Los métodos ágiles [4][5] están centrados en el desarrollo, siendo su objetivo proveer una respuesta rápida a los cambios en los requerimientos, a las personas que componen los equipos y a los problemas que surgen durante el proceso de desarrollo [6][7]. En particular, el proceso de ingeniería de requerimientos en métodos ágiles adopta un enfoque de descubrimiento iterativo [8]. El desarrollo ágil ocurre en un ambiente donde la especificación de especificaciones no ambiguas y completas es imposible o incluso no apropiado [2], por lo que frecuentemente no existe un documento con la especificación de los requerimientos a nivel de sistema y de usuario. Sin embargo, muchas organizaciones que producen software empleando métodos ágiles utilizan pruebas para capturar los requerimientos y los mantienen vinculados al código del software [8]. Este escenario no permite aplicar las prácticas de trazabilidad cómo se lo venía aplicando en los métodos pesados. En consecuencia, es fundamental desarrollar modelos que permitan trazar los requerimientos bajo el enfoque de los métodos ágiles. En la actualidad, una de las metodologías de desarrollo de software ágil más utilizada es Scrum [9][10]. Para poder identificar y definir las posibles trazas en la aplicación de Scrum, se propone en este trabajo un modelo conceptual de Scrum. El modelo debe brindar soporte para responder las siguientes preguntas de competencia: i) ¿Qué eventos originaron un artefacto en particular?; ii) ¿Qué requerimientos guiaron la generación de tal artefacto?; iii) ¿Quiénes son los participantes involucrados en un evento dado? Las respuestas a estar preguntas asistirían las tareas de trazabilidad en proyectos ágiles tales como: Stakeholders con Requerimientos, User Stories con Versiones y Requerimientos con Requerimientos. Los beneficios de estas tareas repercuten directamente en: el análisis del impacto de cambios, conformidad en el producto, obediencia del proceso, responsabilidad del proyecto, reproducibilidad de la línea base y aprendizaje organizacional [2].

La siguiente sección presenta el modelo de Scrum propuesto, el cual se organiza a partir de un conjunto de vistas dados por los conceptos eventos, roles y artefactos que guían a Scrum y las relaciones que existen entre esos conceptos. Luego, en la Sección 3 se presenta un caso de estudio, y por último, en la Sección 4, se presentan las conclusiones del trabajo.

2 Modelo para la Trazabilidad de Scrum

Scrum se define como un “framework” basado en los principios ágiles, utilizado para el desarrollo y gestión de productos complejos, como lo son los productos de software. Puede ser visto como un proceso iterativo e incremental que ayuda a involucrar buenas prácticas ingenieriles dentro de una perspectiva iterativa controlada. En las siguientes secciones se presenta el modelo propuesto centrando cada sección en los constructores esenciales de Scrum: *roles*, *eventos*, *artefactos* y las

reglas que permiten asociar estos conceptos (Fig. 1). Luego, se presentan vistas del modelo para brindar soporte a las respuestas de las preguntas de competencia presentadas en la sección previa.

Fig. 1. Conceptos principales de Scrum.

2.1 Roles

Los roles describen las responsabilidades y niveles de autoridad de individuos o grupo de individuos que participan de manera activa en el proceso. En Scrum se definen los roles *Scrum Team*, *Product Owner*, *Scrum Master*, y *Development Team*. La Fig. 2 incluye tales roles y las relaciones entre los mismos. Un proceso Scrum es llevado a cabo por un equipo de trabajo denominado *Scrum Team* (Fig. 2). Este equipo está conformado por diferentes participantes del proceso de desarrollo, provenientes de la organización desarrolladora del software (*ScrumMaster* y el *DevelopmentTeam*) y de la organización que requiere el software (*ProductOwner*).

El *ProductOwner* (Fig. 2) es un individuo (*Individual* en Fig. 2) miembro de la organización que requiere el software y es el responsable de conducir el proyecto desde la perspectiva del negocio. Es quien debe comunicar una visión clara del producto y definir sus características principales, priorizando los requerimientos del cliente para que el proceso de desarrollo se centre en aquellos requerimientos necesarios para la organización. En todo momento el *ProductOwner* debe contribuir con el equipo (*ScrumTeam* en Fig. 2) para remover todas las dudas que surgen acerca de los requerimientos. Por esta razón, es necesario que el lugar de trabajo del *ProductOwner* esté en el mismo espacio físico que el equipo.

El *ScrumMaster* (Fig. 2) es el encargado de garantizar que los principios de Scrum son aplicados en el proceso de desarrollo. Su función es asegurar que el equipo tenga el conocimiento, las habilidades y la cantidad de personas necesarias para llevar a cabo el trabajo requerido. Usualmente desempeña este rol un individuo que es Scrum Master certificado, un experto en Scrum, para asegurarse que los principios de Scrum sean aplicados.

El *DevelopmentTeam* (Fig. 2) es el equipo de trabajo responsable del desarrollo del producto a entregar, para esto trabaja en conjunto con los distintos stakeholders desde la definición de los requerimientos (*ProductBacklog*, definido en Sección 2.3) hasta la entrega del producto. El tamaño del equipo es una cuestión fundamental para los procesos ágiles, siendo lo común equipos entre 3 y 9 desarrolladores por equipo (relación de agregación en Fig. 2 entre *DevelopmentTeam* y *Developer*). Con más de 9 personas la cantidad de relaciones entre los integrantes aumenta exponencialmente y esto es caótico para la comunicación del equipo [11][12]. El *DevelopmentTeam* puede incluir al *ScrumMaster* y al *ProductOwner* solo en los casos que intervengan en la ejecución de trabajo para el *SprintBacklog* [9] (definido en la Sección 2.3).

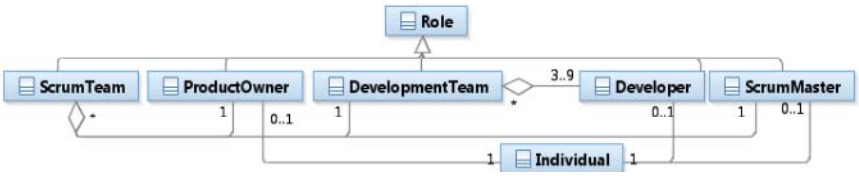


Fig. 2. *Roles* definidos en *Scrum*.

2.2 Eventos

Los eventos (*Event* en Fig. 3) son ocurrencias en el tiempo de un conjunto de reuniones o acciones, utilizados con el objetivo de sincronizar las diferentes etapas por las que atraviesa un proceso Scrum. En La Fig. 3 se representa la relación entre los eventos mediante la asociación *predecesor – successor*. Todo *Event* ocurre en un determinado intervalo de tiempo, representado por *TimeBox* y la asociación *AllocatedOn* en Fig. 3. Un *TimeBox* es una técnica empleada en Scrum para fijar un límite de tiempo en la duración de cada evento o reunión dentro del proceso.

El evento principal en Scrum es un *Sprint*. Un *Sprint* ocurre en un *TimeBox* de 1 mes como máximo y cumple la función de contenedor de eventos. Es utilizado para cumplir los objetivos definidos en él, para generar productos y también para analizar e inspeccionar el proceso. Puede ser analizado como un proyecto debido a que conlleva un esfuerzo temporal para crear un producto y el final se alcanza por el cumplimiento del objetivo o la finalización del proyecto. Un *Sprint* es sucedido por otro, inmediatamente después de que el predecesor es finalizado.

Las distintas reuniones dentro de un *Sprint* se suelen completar de manera secuencial y se estructuran en pequeños *TimeBox*. Esto permite la división del proceso en subconjuntos lógicos que facilitan su dirección, planificación y control. La entidad *Sprint* es llevado a cabo mediante los eventos *SprintPlanning* Meeting, *DailyScrum*, *Sprint Review*, *SprintRetrospective* y la actividad más importante en Scrum que es el desarrollo mismo del producto, *DevelopmentWork* en Fig. 3.

El *SprintPlanning* es una reunión donde se plantea como objetivo elicitar los requerimientos del cliente y la planificación de qué artefactos serán entregados y cómo se construirán. La entidad *SprintPlanning* (Fig. 3) posee un *TimeBox* de ocho horas para un *Sprint* de un mes y tiene una duración proporcionalmente menor para *Sprint* más cortos. Está compuesta por dos partes (*FirstPart* y *SecondPart* en Fig. 3), cada una con la mitad de la duración de toda la reunión *SprintPlanning*. En la reunión *FirstPart* (Fig. 3) se desarrolla el qué se hará durante el *Spring*. El objetivo de esta reunión es que el *DevelopmentTeam* entienda en detalle los requerimientos del usuario. Con esto, para finalizar esta etapa deciden cuales requerimientos están en condiciones para ser desarrollados. En la reunión *SecondPart* (Fig. 3) se desarrolla el "cómo" se obtendrán los mismos. El objetivo de esta actividad es que el *DevelopmentTeam* determine qué interfaces necesitará implementar, qué arquitectura deberá crear y que tablas o componentes requerirán ser actualizados o desarrollados.

El *DailyScrum* es una reunión corta donde el *DevelopmentTeam* coordina y planea su siguiente día de trabajo reportando avances y dificultades. Un *DailyScrum* (Fig. 3)

ocurre en un *TimeBox* de 15 minutos realizado cada día de trabajo. Este evento es seguido por el *DevelopmentWork***,** la parte más importante del *Sprint* donde se implementa la solución. Además, en esta etapa se realizan las pruebas y al finalizar se entrega el incremento logrado en el producto (*ProductIncrement*, descripto en Sección 2.3). Luego del *DevelopmentWork* sucede una *SprintReview*. Esta es una reunión en la que se inspecciona el incremento de producto (*ProductIncrement*) creado y se adapta el *ProductBacklog* (explicados en Sección 2.3) para el siguiente *Sprint*. Como resultado del *SprintReview* se obtiene información de entrada para el siguiente *SprintPlanning*. Generalmente un *SprintReview* sucede en un *TimeBox* de cuatro horas para una *Sprint* de un mes.

La otra reunión de inspección posible en un *Spring* es la *SprintRetrospective*, la cual se desarrolla luego de la *SprintReview* y antes de la próxima reunión de *SprintPlanning* (ver Fig. 3). Este evento tiene como objetivo revisar cómo fue realizado el último *Sprint* (*Roles*, *Eventos*, y *Artefactos*). En ella se crea un plan de mejoras que serán aplicadas durante la siguiente iteración. Usualmente sucede en un *TimeBox* de tres horas para una *Sprint* de un mes.

Durante un Sprint puede surgir un evento denominado *Grooming* (Fig. 3). El *Grooming* es una especificación (división) de una tarea en tareas más pequeñas.

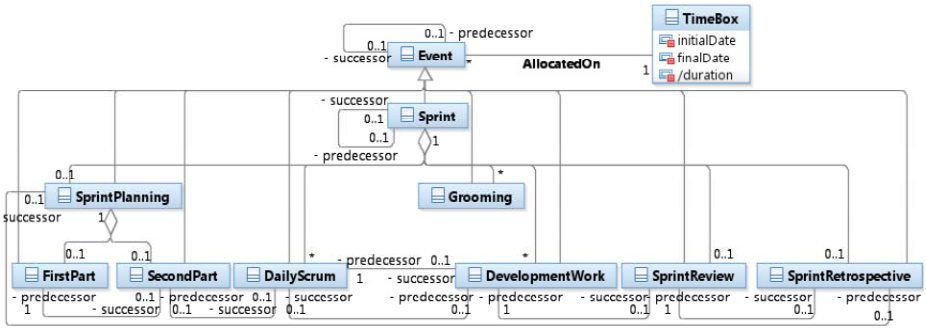


Fig. 3. Representación de eventos definidos en Scrum.

2.3 Artefactos

Un artefacto es la descripción de un producto de trabajo. Los artefactos pueden estar compuestos por otros artefactos y son productos de trabajos concretos consumidos, producidos o modificados por los distintos eventos del proceso [12]. En la Fig. 4 se representan los conceptos vinculados a los artefactos del proceso.

La generación del producto principal está guiada por un conjunto ordenado de requerimientos denominado *ProductBacklog*. El *product backlog* contiene cada requerimiento (*ProductBacklogItem*) que podría ser tratado en el desarrollo de un producto (*Product*), como así también en cada incremento del producto (*ProductIncrement*). Habitualmente la lista de *ProductBacklogItem* se encuentra ordenada por valor, riesgo, prioridad y necesidad. Asimismo, el incremento del producto sirve luego como retroalimentación, permitiendo la evaluación del

incremento y la generación de nuevos requerimientos. De esta manera se representa un *Feedback* entre *ProductIncrement* y *ProductBacklogItem*. Los *ProductBacklogItem* son requerimientos que plantean necesidades, deseos o expectativas que el *ScrumTeam* quiere entregar en el futuro. Estos ítems son parte del *ProductBacklog* y a su vez pueden ser refinados (*RefinedOn*) en nuevos elementos durante un evento *Grooming* (Fig. 3).

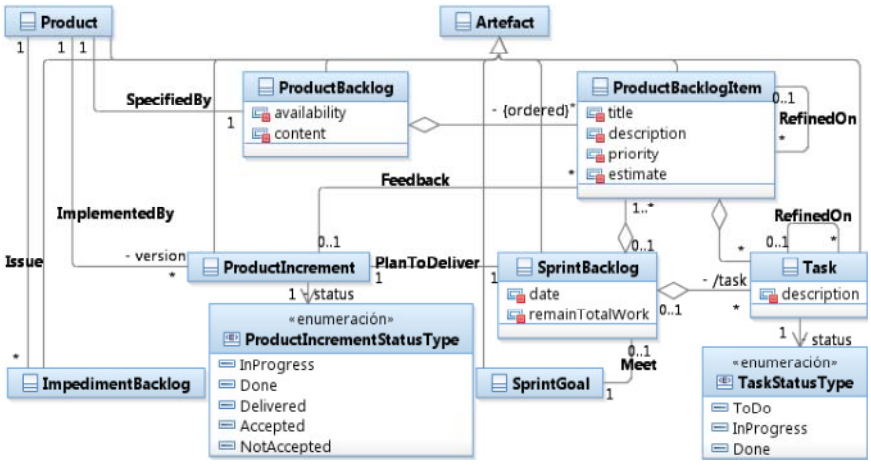


Fig. 4. Representación de Artefactos de Scrum.

El *SprintBacklog* es una lista de tareas (*Task*) que indica las actividades que se tienen que realizar durante el *Sprint* y cuál es la situación actual del equipo de desarrollo. El *SprintBacklog* reúne al conjunto de *ProductBacklogItem* seleccionados para ser trabajados durante un *Sprint* y con ello poder cumplir el *SprintGoal*, además comprende su planificación para entregar el *ProductIncrement* (relación *PlanToDeli*ver).

El *ProductIncrement* es el resultado de la implementación de todos los *ProductBacklogItem* especificados por el *SprintBacklog*, durante un *Sprint.* Al finalizar el Sprint el *ScrumTeam* entrega esta versión parcial del producto de software. Este entregable tiene una propiedad *status* la cual representa el estado en que se encuentra (*ProductIncrementStatusType*): *InProgress*, *Done*, *Delivered*, *Accepted*, o *NotAccepted*.

Los distintos ítems del *ProductBacklog* que componen el *SprintBacklog* se descomponen en un conjunto de tareas (*Task* en Fig. 4) que indican lo que el equipo de desarrollo debe realmente hacer para lograr el *SprintGoal*. Durante el proceso de desarrollo cada tarea puede estar en el *status* (*TaskStatusType* en Fig. 4): *ToDo*, *InProgress*, o *Done*. *ToDo* representa una tarea del *SprintBacklog* que está sin realizarse. Si permanece más de un día de trabajo en este estado es fragmentada en tareas más pequeñas [13] (relación *RefinedOn*). *InProgress* es cuando la tarea fue comenzada pero aun no finalizada. *Done* representa que la tarea fue finalizada.

Las distintas dificultades presentes en el proceso se incorporan en el *ImpedimentBacklog* , una lista de dificultades, impedimentos u obstáculos que limitan el rendimiento del *DevelopmentTeam*. El *ScrumMaster* tiene la responsabilidad de removerlos lo antes posible.

2.4 ¿Qué Eventos generaron un Artefacto en particular?

Los eventos son secuenciales y la conclusión de un evento finaliza con la entrega de un artefacto producido (relación *Output* en Fig. 1). Así también la finalización de un evento representa un punto a ser evaluado y adaptado. Cada artefacto (Fig. 4) es una herramienta o un producto de trabajo intermedio que permite llevar a cabo el trabajo en entornos difíciles. Cada *Sprint* tiene una visión única que la hace diferente de cualquier otra que pudiera sucederla e intenta alcanzar un cierto objetivo (*SprintGoal* en Fig. 5). En la Fig. 5 se representan las entradas y salidas de un *Sprint*. El modelo es extensible a los distinots eventos que componen el *Sprint*. En Fig. 5 se incluye a modo de ejemplo los eventos *FirstPart* del *SprintPlanning* y *DailyScrum*.

Durante la reunión *FirsPart* del *SprintPlanning* se define el objetivo del *Sprint*, *SprintGoal*, y los *ProductBacklogItem* que serán desarrollados durante la *Sprint* en curso, representados por la relación *Output* de *FirstPart* en Fig. 5.

En la realización del *DailyScrum* se coordina y planea el siguiente día de trabajo reportando avances y dificultades, se seleccionan los elementos en los que se trabajará, representado por la relación *Satisfy ProductBacklogItem* (Fig. 5), y se intenta resolver las dificultades presentes (relación *Remove ImpedimentBacklog* en Fig. 5). Las relaciones *Satisfy* y *Remove* especializan la relación *Output* definida entre *Event* y *Artefact* en Fig. 1.

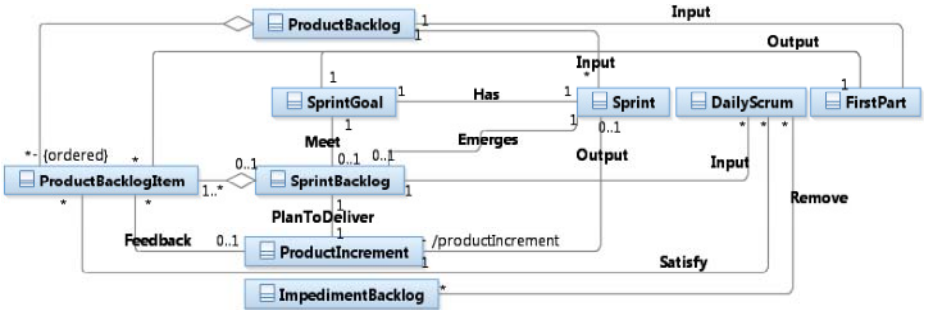


Fig. 5. Representación de Artefactos vinculados a los Eventos *Sprint*, *FirstPart*, y *DailyScrum*.

2.5 ¿Qué Requerimientos Guiaron la Generación de cierto Artefacto?

Los eventos representados en el modelo propuesto junto a las relaciones con los distintos artefactos de un proceso Scrum permiten obtener información acerca de qué requerimientos fueron considerados en la generación de los distintos artefactos, principalmente de los incrementos del producto principal de software (*ProductIncrement* en Fig. 5). En Scrum cada *Sprint* posee un objetivo (*SprintGoal*)

el cual es materializado por un *SprintBacklog*, el cual reúne un conjunto de requerimientos detallados en el *ProductBacklogItem*. La representación explícita de estas relaciones (*Has*, *Meet*, y *Emerges* en Fig. 5) permite conocer los requerimientos que guiaron los distintos *Sprint* del proceso de desarrollo, en los cuales se generaron los distintos incrementos del producto (*ProductIncrement* en Fig. 5).

2.6 ¿Quiénes son los Participantes Involucrados en un Evento dado?

Como fue mencionado previamente, los eventos son realizados (*Perform* en Fig. 1) por uno o más individuos (*Individual* en Fig. 2) ejerciendo ciertos *Roles*. La relación *Perform* explicitada en la Fig. 1 es especializada para representar las distintas formas de participación en el proceso. Por limitaciones de espacio en esta sección sólo se presenta la especialización para el evento *Sprint*. La Fig. 6 incluye dicho evento y los distintos roles que participan del mismo.

Una de las principales responsabilidades del *ProductOwner* es definir qué se debe realizar en el *Sprint* (relación *Define* en Fig. 6). Selecciona los cinco *ProductBacklogItems* más importantes, son asignados al *Sprint*, y se asegura que el equipo pueda desarrollarlas en el *Sprint*. La otra responsabilidad que posee es revisar y aprobar el trabajo en el *Sprint* diciendo si lo que fue creado o "hecho" satisface el objetivo del *Sprint* (relación *Aprove* en Fig. 6).

La función primordial del *ScrumMaster* es asegurar que las políticas de la organización sean adoptadas por el *DevelopmentTeam* y de revisar cómo fue realizado el último Sprint, los *Roles*, *Eventos*, y *Artefactos* que lo componen.

El *ScrumMaster* debe trabajar junto al *ProductOwner* y ambos tienen la responsabilidad de intermediar para lograr un entendimiento o negociación sobre algún problema en los tiempos de trabajo o agregar *ProductBacklogItem* al *Sprint*.

El *DevelopmentTeam* participa dentro de un *Spring* realizando todo el *DevelopmentWork*, toma los requerimientos, los completa y satisface, creando un producto de software. Realiza las tareas de análisis, diseño, desarrollo, testing, documentación y mantenimiento. El equipo decide que será realizado y por cuál actor dentro del equipo. Se auto organiza para planear cómo va a realizar el trabajo dentro del *Sprint*, esto es un punto crítico de Scrum.

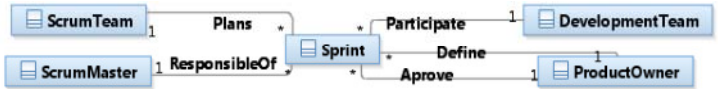


Fig. 6. Representación de los *Roles* participando en un *Sprint*.

3 Caso de Estudio

El modelo presentado en la Sección 2 se lo aplica a un proyecto de sistema de comercio electrónico. El *ProductOwner* presenta los *ProductBacklogItem* que considera que deben ser estimados. El *ScrumTeam* divide los ítems del *Backlog* en ítems más pequeños (*Grooming*), y redimensiona el *ProductBacklog* con los nuevos

elementos; por ejemplo, en la Fig. 7 se ilustra el *ProductBacklog*, donde el ítem (*ProductBacklogItem*) *Checkout* fue refinado (*RefinedOn*) en los ítems *Selección Formas de Pago* y *Selección Medio Envío.*

En el evento *FirstPart* del *Sprint* 1 el *ProductOwner* ordena el *ProductBacklog* y define cuál será el *SprintGoal,* en este caso *Diseño Frontend*, *Administrar Productos*, y *Diseño Base de Datos.* En la *SecondPart,* el *DevelopmentTeam* realiza bosquejos para crear o actualizar las interfaces, la arquitectura y/o los componentes. Además, define los *ProductBacklogItem* a desarrollar en el *Sprint1* (*SprintBacklog* en Fig.8).

La Fig. 8 presenta el *SprintBacklog* junto a información de los *ProductBacklogItems* y *Tasks* que lo componen. La figura muestra el backlog una vez iniciada la tarea de "Diseñar Header, Body, Footer" por el *DevelopmentTeam*, donde el *Status* de la tarea fue pasado de *ToDo* a *InProgress*.

	Title	ID	Owner	Priority
	Sprint1			
	Diseño Frontend	S-01010		High
	Administrar Productos	S-01004		High
	Diseño Base de Datos	S-01012		Medium
	Sprint2			
Checkout	Seleccion Formas de Pago	S-01003		High
	Seleccion Medio Envío	S-01013		High
	Implementación del carro	S-01008		High
	Administrar Cuenta de Usuario	S-01002		Low

Title	ID	Owner	Priority
Sprint3			
Administrar Pedidos	S-01007		Medium
Ingreso al Sistema	S-01001		Low
Administrar Usuarios	S-01005		Low
Diseño Backend	S-01011		Low

Fig. 7. *ProductBacklog* del caso de estudio.

PBI	RemainTotalWork	Title	PBI-ID	TaskID	Estimate	Priority	Status
Diseño Frontend	50,00%	Diseñar Header, Body, Footer	S-01010	T-0001		High	InProgress
		Crear el Estilo	S-01010	T-0002			Done
Administrar Productos	100,00%	Crear Productos	S-01004	T-0003		High	InProgress
		Actualizarr Productos	S-01004	T-0004			ToDo
		Eliminar Productos	S-01004	T-0005			ToDo
Diseño Base de Datos	100,00%	Crear Tablas y Atributos	S-01012	T-0006		Medium	InProgress

Fig. 8. *SprintBacklog* de la iteración número 1.

Una vez concluido el *Sprint 1,* el *DevelopmentTeam* entrega los *ProductBacklogItem* con *status Done* al *ProductOwner.* Este último es el encargado de inspeccionar si cumple con la definición de producto terminado, aprobando o rechazándolo. Una vez aprobado el *ProductBacklogItem*, el *ScrumMaster* organiza el *Sprint Review*. Además toma nota del *Feedback* del encuentro del *ProductOwner* con el usuario final, donde es presentado el *ProductIncrement.* Si este es aprobado el *ScrumTeam* libera *e-Commerce versión 0.1*.

Durante el *SprintRetrospective* el *ScrumTeam* evalúa el *Sprint1,* registrando los logros alcanzados, las dificultades que surgieron y las sugerencias de cómo mejorar el proceso. Una vez finalizado esto, el *ScrumTeam* se da inicio al *Sprint2*.

Para poder validar la aplicación del modelo propuesto al ejemplo presentado se utilizó la herramienta USE (UML Specification Environment) la cual permite validar y verificar especificaciones consistentes de diagramas de clases UML y realizar consultas en OCL [14]. La validación se realizó generando las instancias del modelo propuesto siguiendo la ejecución previamente enunciada del caso de estudio. A partir

del mismo, fue posible conocer los requerimientos (instancias de *ProductBacklogItem*) considerados en el incremento de producto *e-Commerce versión 0.1*, como así también los miembros del *Scrum Team* que participaron en el *Sprint 1* que generó tal artefacto.

4 Conclusiones

En este trabajo se propone un modelo conceptual de Scrum para representar los eventos que lo componen junto a los artefactos generados. En dicho modelo se establecieron diferentes conceptos que especializan los conceptos básicos de Scrum y que pueden ser utilizados para asistir en tareas de trazabilidad en proyectos ágiles tales como: Stakeholders con Requerimientos, User Stories con Versiones y Requerimientos con Requerimientos. Los trabajos futuros utilizarán el modelo conceptual propuesto para formalizar el soporte a tales tareas.

5 Referencias

1. IEEE Standard Glossary of Software Engineering Terminology. IEEE Std 610.12-1990 1–84 (1990)
2. Espinoza, A., Garbajosa, J.: A study to support agile methods more effectively through traceability. Innovations in Systems and Software Engineering 7, 53–69 (2011)
3. Pikkarainen, M., Passoja, U.: An Approach for Assessing Suitability of Agile Solutions: A Case Study. In: 6th International Conference on Extreme Programming and Agile Processes in Software Engineering (2005)
4. Agile Manifesto, http://www.agilemanifesto.org (2001)
5. Williams, L.: What agile teams think of agile principles. ACM Communications 55, 71–76 (2012)
6. Hunt, J.: Agile Software Construction. Springer (2005)
7. Dingsøyr, T., Nerur, S., Balijepally, V., Moe, N.: A decade of agile methodologies: Towards explaining agile software development. Journal of Systems and Software 85, 1213–1221 (2012)
8. Cao, L., Ramesh, B.: Agile Requirements Engineering Practices: An Empirical Study. IEEE Software 25, 60–67 (2008)
9. Sutherland, J., Schwaber, K.: The Scrum Guide. The Definitive Guide to Scrum: The Rules of the Game (2011)
10. Schwaber, K.: SCRUM Development Process. In: OOPSLA'95 Workshop on Business Object Design and Implementation (1995)
11. Brooks , F.: Mythical Man-Month, The: Essays on Software Engineering, 2nd ed. Addison-Wesley Professional (1996)
12. OMG: Software & Systems Process Engineering Metamodel Specification (SPEM) Version 2.0 (2008)
13. Glogler, B.: Scrum Checklist 2012. bor!sgloger Wien. Baden-Baden. (2012)
14. P. Ziemann, M. Gogolla: Validating OCL Specifications with the USE Tool – An Example Based on the BART Case Study. Electronic Notes in Theoretical Computer Science 80, 157-169 (2003)

Evaluación de variantes en modelo destinado a anticipar la conveniencia de trazar proyectos de software

Juan Giró, Juan Vázquez, Brenda Meloni y Leticia Constable

Departamento de Ingeniería en Sistemas de Información
Facultad Regional Córdoba, Universidad Tecnológica Nacional
Maestro López esq. Cruz Roja Argentina, Ciudad de Córdoba
{juanfgiro, jcjvazquez, bemeloni, leticiaconstable}@gmail.com

Resumen. La escasez de evidencias de que los progresos en el campo de la trazabilidad sean efectivamente aprovechados por la industria del software ha estimulado el desarrollo de modelos conducentes a un mejor conocimiento del problema y poder anticipar los resultados esperables en proyectos. Para ello fue necesario identificar los factores de mayor impacto en el éxito de los procesos de trazabilidad y proponer modelos que permitan hacer predicciones a partir de esos factores. En este trabajo se evalúan los resultados de introducir variantes en las métricas asociadas a esos factores con el fin de posibilitar la selección de las más convenientes para el mejor desempeño del modelo de predicción. Se utiliza para ello el Análisis ROC, que a pesar de sus ventajas ha tenido hasta el momento poca difusión en la ingeniería de software.

Keywords: ingeniería de software, análisis ROC, trazabilidad de requerimientos.

1 Introducción

Las evidencias de que los progresos en el campo de la trazabilidad de requerimientos de proyectos de desarrollo no llegan a ser efectivamente aplicadas en la industria del software [1] condujeron a la necesidad de entender mejor el problema y sus causas. Revisando las experiencias desfavorables se comprobó que pueden ser reunidas en tres grupos: *i*) las que fueron prematuramente abandonadas o no cubrieron las expectativas desde un punto de vista técnico, *ii*) las técnicamente exitosas con un costo de implementación mayor que el beneficio obtenido y *iii*) las que con un elevado costo condujeron a un resultado pobre o nulo, es decir una combinación de las dos primeras. Lo expuesto resulta sorprendente ya que en la actualidad es unánime el reconocimiento de la trascendencia e importancia de la trazabilidad de requerimientos como soporte de los procesos de desarrollo de software [2], habiendo sido incorporada en todas las normas y modelos de desarrollo vigentes.

Antes de continuar es necesario enfatizar que al hablarse de trazabilidad de requerimientos en proyectos de desarrollo de software se está haciendo referencia a una gestión que vincula las numerosas etapas de sus ciclos de vida, asegurando el éxito del proyecto, brindando la necesaria garantía de coherencia, completitud y

corrección al software producido y posibilitando su eficaz mantenimiento correctivo y preventivo en el resto de su vida útil.

Al analizarse las líneas de estudio en el campo de la trazabilidad, se comprueba que en su mayor parte están orientadas a desarrollar nuevas metodologías y herramientas, habiendo un esfuerzo mucho menor destinado a estudiar el resultado de la aplicación de las mismas en la industria y las causas de las dificultades ya señaladas. Además, los escasos documentos [1][3] destinados a analizar el origen de las dificultades de la trazabilidad abordan el problema en forma cualitativa y en la mayoría de las veces el enfoque es demasiado general.

Surgió así la presunción de que no es fortuito que ciertos proyectos puedan ser exitosamente trazados y otros no, por lo que debe haber una combinación de condiciones objetivas que conducen a uno u otro resultado, y esta idea llevó a plantear la hipótesis de que *existen factores que condicionan el éxito de los procesos de trazabilidad y que es factible identificarlos*.

La comprobación de la hipótesis enunciada orienta la actividad cumplida en el proyecto "Aseguramiento de la Trazabilidad en Proyectos de Desarrollo de Sistemas de Software" [4], y en este marco se propusieron factores y modelos destinados a anticipar los resultados de procesos de trazabilidad, que vienen siendo progresivamente mejorados [5][6][7].

En este trabajo se estudian variantes a las métricas propuestas para la evaluación de los ya mencionados factores con la finalidad de identificar las más convenientes. La organización del documento es la siguiente: en la sección 2 se resumen las características del modelo estudiado, los factores de trazabilidad elegidos y las variantes propuestas para sus métricas, en la sección 3 se analiza el impacto de las nuevas métricas en el desempeño del modelo, discutiendo los resultados obtenidos, y en la Sección 4 se presentan las conclusiones de este trabajo y actividades futuras.

2 Modelo de trazabilidad de proyectos, sus factores y métricas

Se reconocen tres entidades principales que están estrechamente relacionadas entre sí, que son: *a*) el propio *producto software*, *b*) el *proyecto*, que responde a cierto modelo de proceso y ampara la construcción del producto y *c*) la *organización*, que constituye el escenario en el que el proyecto es desarrollado. Por lo tanto, se anticipa que los factores buscados estarán asociados a dimensiones de estas tres entidades.

También se establecieron criterios con respecto a la selección de los factores, la forma en que éstos son evaluados y su interpretación con respecto al problema tratado, que son los siguientes: *a*) deben ser *cuantificables*, *b*) se les aplicará un multiplicador de escala para expresarlos en el *intervalo cero a cinco*, *c*) los valores *crecientes* contribuyen más favorablemente a la trazabilidad de un proyecto y *d*) deben ser *ortogonales entre sí*.

Se definen así, agrupados por entidades, los ocho factores propuestos [6][7] para predecir la conveniencia de realizar la trazabilidad de requerimientos en proyectos de desarrollo de software, que son resumidos a continuación en las Tabla 1 y Tabla 2:

Tabla 1: Definición de entidades, factores, variables asociadas, descripciones y variables de referencia

Entidad	Factor	Variable	Descripción y variable de referencia ***Vr***	Intervalo de ***Vr***
Producto	Tamaño	***t***	Puntos de Función ***PF***	100 - 1000
	Vigencia	***v***	Vida útil ***VU*** [años]	0,5 - 10
	Reutilización	***r***	Futura reutilización ***RE*** [%]	0 - 80
	Confiabilidad	***c***	Indicador confiabilidad ***CO*** (*)	0 - 5
Proyecto	Plazo	***p***	Duración proyecto ***DP*** [años]	0 - 5
	Equipo	***e***	Efectividad del equipo ***EF*** (**)	1 - 5
Organización	Madurez	***m***	Nivel de madurez ***CMMI***	1 - 5
	Dependencia	***d***	Nivel de autonomía ***NA*** (***)	1 - 5

Tablas 2: Definición de los indicadores ***CO*** (*), ***EF*** (**), Nivel de Madurez ***CMMI*** y ***NA*** (***)

Confiabilidad ***CO***	
0	No importante
1	Baja
2	Media
3	Alta
4	Muy alta
5	Absoluta

Efectividad ***EF***	
1	Pobre
2	Baja
3	Media
4	Alta
5	Muy alta

Madurez ***CMMI***	
1	Inicial
2	Gestionado
3	Definido
4	Cuant.Gestion.
5	Optimizado

Autonomía ***NA***	
1	Independiente
2	Normas propias
3	Normas clientes
4	Nor. casa matriz
5	Combina 3 y 4

En las referencias [6] y [7] se dispone de un detalle del alcance y justificación de los factores propuestos y las métricas adoptadas para el modelo presentado, que en lo sucesivo se denominará Modelo "A".

A título ilustrativo se muestra en la Figura 1 la representación de los ocho factores recurriendo a un diagrama de radar, donde los polígonos ejemplifican: *a*) un proyecto trazable, *b*) un proyecto no trazable, *c*) un proyecto atípico y *d*) la "zona gris" que representa la frontera discriminante entre ambos casos.

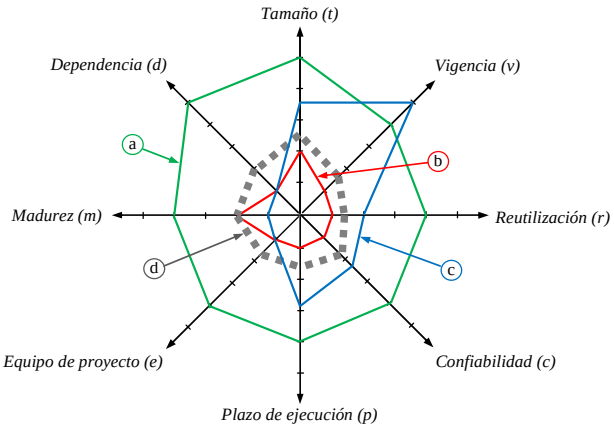


Figura 1: Representación de la trazabilidad de proyectos en un diagrama de radar

Con el Modelo “A” se propuso la asignación de una variable a cada una de las tres entidades involucradas: el *producto* (η_1), el *proyecto* (η_2) y la *organización* (η_3), donde se tuvo en cuenta la ortogonalidad entre los factores al calcular el módulo resultante de cada una de las tres variables. Esto permite reducir la dimensión del problema y facilita la visualización de las poblaciones de datos, presentándose en la Tabla 3 las expresiones de las tres variables η_1, η_2 y η_3.

Tabla 3: Reducción del problema a tres dimensiones

Entidad	Variable	Expresión para su evaluación
Producto	η_1	$\eta_1 = \sqrt{\left(\boldsymbol{t}^2 + \boldsymbol{v}^2 + \boldsymbol{r}^2 + \boldsymbol{c}^2\right)}$
Proyecto	η_2	$\eta_2 = \sqrt{\left(\boldsymbol{p}^2 + \boldsymbol{e}^2\right)}$
Organización	η_3	$\eta_3 = \sqrt{\left(\boldsymbol{m}^2 + \boldsymbol{d}^2\right)}$

A partir de esta reducción de dimensiones surgió la idea de utilizar el módulo de la resultante de los ocho factores, en adelante denominado “$\boldsymbol{\rho}$”, como parámetro representativo o “indicador” de cada caso considerado:

$$\boldsymbol{\rho} = \sqrt{\left(\eta_1^2 + \eta_2^2 + \eta_3^2\right)} = \sqrt{\left(\boldsymbol{t}^2 + \boldsymbol{v}^2 + \boldsymbol{r}^2 + \boldsymbol{c}^2 + \boldsymbol{p}^2 + \boldsymbol{e}^2 + \boldsymbol{m}^2 + \boldsymbol{d}^2\right)} \qquad (1)$$

El indicador $\boldsymbol{\rho}$ representa el radio de una fracción de casquete esférico en el espacio de tres dimensiones, o en el hiperespacio de ocho, y el objetivo es determinar el valor más apropiado de $\boldsymbol{\rho}$ para que queden separadas de la mejor forma las poblaciones de los proyectos no trazables de los trazables. En la Figura 2.a se muestra la población de datos tomada como “caso de estudio” sobre el sistema de ejes cartesiano η_1, η_2 y η_3 y en la Figura 2.b la distribución de estos datos en función de $\boldsymbol{\rho}$.

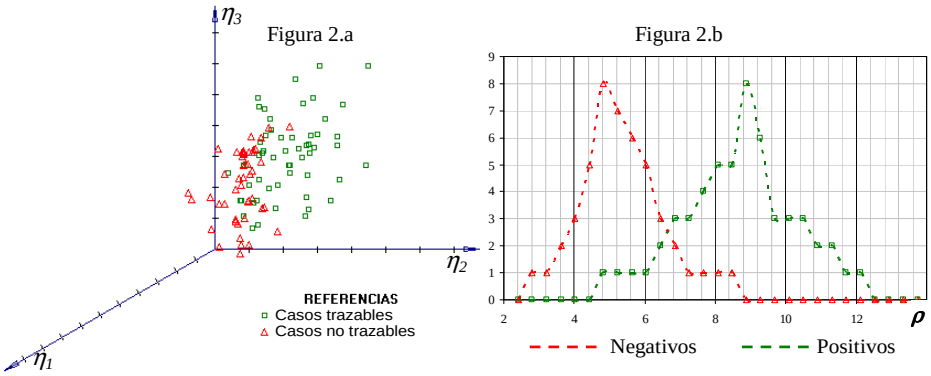


Figura 2: a) Representación de la población del lote de datos del caso de estudio y b) Curvas de distribución de la población de datos en función de “$\boldsymbol{\rho}$”

Si bien el modelo “A” propuesto demostró un buen desempeño [7], se presentó el interrogante sobre si las métricas utilizadas para cuantificar los ocho factores eran las

más apropiadas. Expresado en otros términos, ¿es posible obtenerse una mejor discriminación de los casos positivos y negativos si se utilizan otras métricas en la definición de ρ? De ser así las poblaciones mostradas en la Figura 2.b quedarán más separadas y el modelo será más efectivo para clasificarlas.

Para responder a este interrogante se incorporaron cuatro variantes al Modelo "A" originalmente utilizado, que son resumidas en la Tabla 4 a continuación:

Tabla 4: Definición de las métricas de variables en el modelo "A" y sus variantes

Varia-ble	Modelo A (original)	Variante B (lineal)	Variante C (lineal modif.)	Variante D (semi sigm.)	Variante E (sigmoidal)
t	5****PF*** / 1000	5****PF*** / 1000	5****PF*** / 1000	Sg(5****PF***/1000)	Sg(5****PF***/1000)
v	5****VU*** / 10	5****VU*** / 10	5****VU*** / 10	Sg(5****VU***/ 10)	Sg(5****VU***/ 10)
r	Ver (*)	1 + ***RE***/20	***RE***/16	Sg(1+ ***RE***/20)	Sg(***RE***/16)
c	1 + 0,16****CO***^{2}	1 + 0,8****CO***	***CO***	Sg(1+0,8****CO***)	Sg(***CO***)
p	***DP***	***DP***	***DP***	***DP***	Sg(***DP***)
e	***EF***	***EF***	***EF***	***EF***	Sg(***EF***)
m	***CMMI***	***CMMI***	***CMMI***	***CMMI***	Sg(***CMMI***)
d	***NA***	***NA***	***NA***	***NA***	Sg(***NA***)

(*) $\boldsymbol{r} = 1 + 0{,}025*(\boldsymbol{RE} + 0{,}0125*\boldsymbol{RE}^2)$

En las variantes "D" y "E" se adopta una expresión sigmoidal para cuantificar las variables de los factores de trazabilidad. El objeto de la función sigmoidal es polarizar los resultados hacia los extremos del intervalo. Para implementar esta expresión se incorpora un corrimiento del origen y un factor de amplificación para brindar resultados en el intervalo [0,5] para valores del argumento $0 \leq x \leq 5$, tal como se representa en la Ec.2:

$$Sg(x) = 5 / (1 + \exp(-2*(x-2{,}5)) \qquad (2)$$

Analizando las métricas presentadas en la Tabla 2 se comprueba que:

Variable *t*: Para el modelos "A" y variantes "B" y "C" se propone una misma fórmula lineal, mientras que en las variantes "D" y "E" el resultado de la fórmula lineal es afectado de la expresión sigmoidal (Ec. 2).

Variable *v*: Se adopta el mismo criterio de la variable anterior: una fórmula lineal para los primeros tres casos y una corrección sigmoidal (Ec. 2) para los dos últimos.

Variable *r*: Se propone una fórmula polinomial para el Modelo "A", una expresión lineal que brinda resultados en intervalo [1..5] en la variante "B", una expresión lineal con resultados en el intervalo [0..5] en la variante "C", la corrección sigmoidal de la variante "B" es asignada a la variante "D" y la corrección sigmoidal de "C" es asignada a la variante "E". Con la finalidad se facilitar la interpretación del efecto esperado con las diferentes expresiones se las representa en la Figura 3 en función del porcentaje de reutilización ***RE*** en el intervalo 0 – 80%.

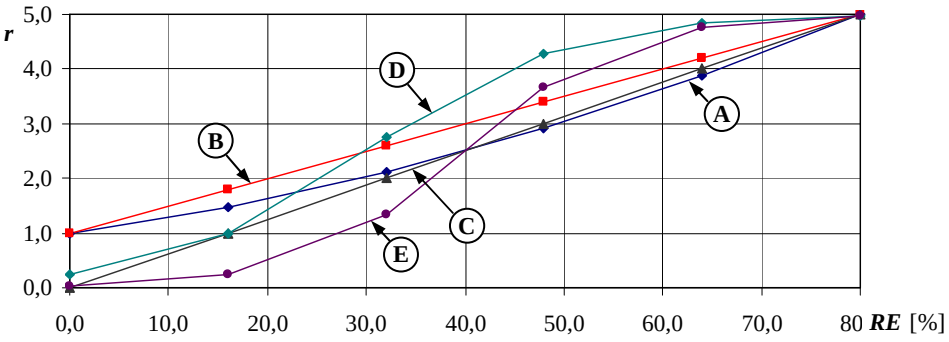


Figura 3: Evolución de la variable ***r*** en función del porcentaje de reutilización ***RE*** según las expresiones propuestas para las métricas de las diferentes variantes.

Como ejemplo se consideran dos puntos; RE = 16 % y RE = 64 %, y en la Tabla 5 se muestran los valores de ***r*** obtenidos en cada caso, observándose que el caso E presenta una polarización antisimétrica y el D una polarización asimétrica.

Tabla 5: Valores de ***r*** obtenidos con diferentes métricas en dos puntos (***RE*** = 16 y 64%)

RE = 16 %		***RE*** = 64 %	
Mod.	***r***	**Mod.**	***r***
A	1,48	A	3,88
B	1,80	B	4,20
C	1,00	C	4,00
D	0,98	D	4,84
E	0,24	E	4,76

Variable *c*: En los diferentes modelos las fórmulas propuestas son similares a las del caso anterior: "A" polinomial, "B" lineal en el intervalo [1..5], "C" lineal en el intervalo [0..5] y las correcciones sigmoidales de "B" y "C" en las dos últimas. Se representan las expresiones en la Figura 4.

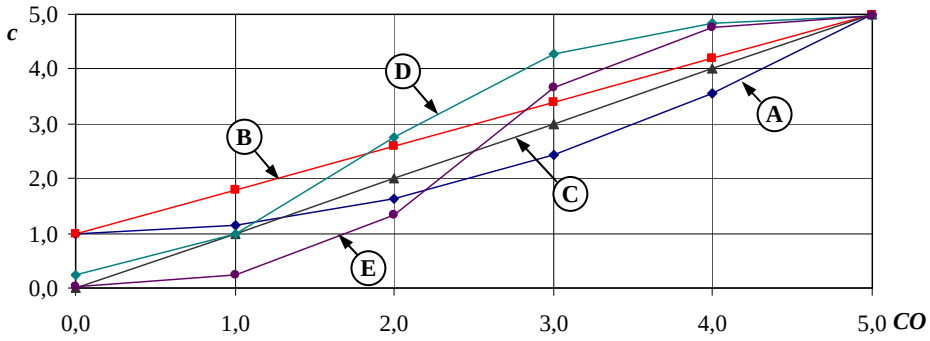


Figura 4: Evolución de la variable ***c*** en función de la confiabilidad CO según las expresiones propuestas para las métricas de las diferentes variantes.

Variable *p*: Para el modelos "A" y variantes "B", "C" y "D" se propone la asignación

directa del plazo de proyecto ***DP*** a ***p***, mientras que en las variante "E" se asigna el resultado aplicar la expresión sigmoidal al plazo ***DP*** (Ec. 2).

Variable *e*: Se adopta el mismo criterio que para ***p***, en este caso para ***EF***.

Variable *m*: Se adopta el mismo criterio que para ***p***, en este caso para el nivel de madurez ***CMMI***.

Variable *d*: Se adopta el mismo criterio que para ***p***, en este caso para el nivel de autonomía ***NA***.

Una vez establecidas las variantes a ser consideradas, el paso siguiente fue definir la herramienta de comparación a efectos de identificar el modelo más conveniente. Para este caso resulta recomendable el análisis ROC [8], que es una técnica destinada a evaluar clasificadores dicotómicos y que recientemente ha experimentado gran difusión en campos muy variados tales como la bioingeniería, aprendizaje automático y minería de datos, aunque todavía es poco usado en la ingeniería de software. El Análisis ROC permite: *i*) poder elegir objetivamente el mejor entre varios modelos de clasificación y *ii*) optimizar la sintonía del modelo elegido. En este caso se lo aplica para lo primero.

Si se considera poblaciones de datos positivos y negativos, tales como las representados en la Figura 2.b, al definirse un valor de corte para el indicador $\boldsymbol{\rho}$ (denominado $\boldsymbol{\rho}_c$) quedan inmediatamente establecidos cuatro agrupamiento de los datos que son representados en la Tabla 6. Estos agrupamientos dan lugar a la definición de parámetros que son la base del análisis ROC, según se muestra a continuación.

Tabla 6: Resultados obtenidos con un clasificador y definición de parámetros

Modelo Clasificador		Condición real	
		Positivos	**Negativos**
Resultados con valor de corte $\boldsymbol{\rho}_c$	**Positivos**	**Verdaderos positivos (*VP*)**	**Falsos positivos (*FP*)**
	Negativos	**Falsos Negativos (*FN*)**	**Verdaderos negativos (*VN*)**

$$\boldsymbol{Sensibilidad = FVP = \frac{VP}{VP + FN}} \quad (3)$$

$$\boldsymbol{Especificidad = FVN = \frac{VN}{VN + FP}} \quad (4)$$

$$\boldsymbol{Especificidad = FVN = 1 - FFP} \quad (5)$$

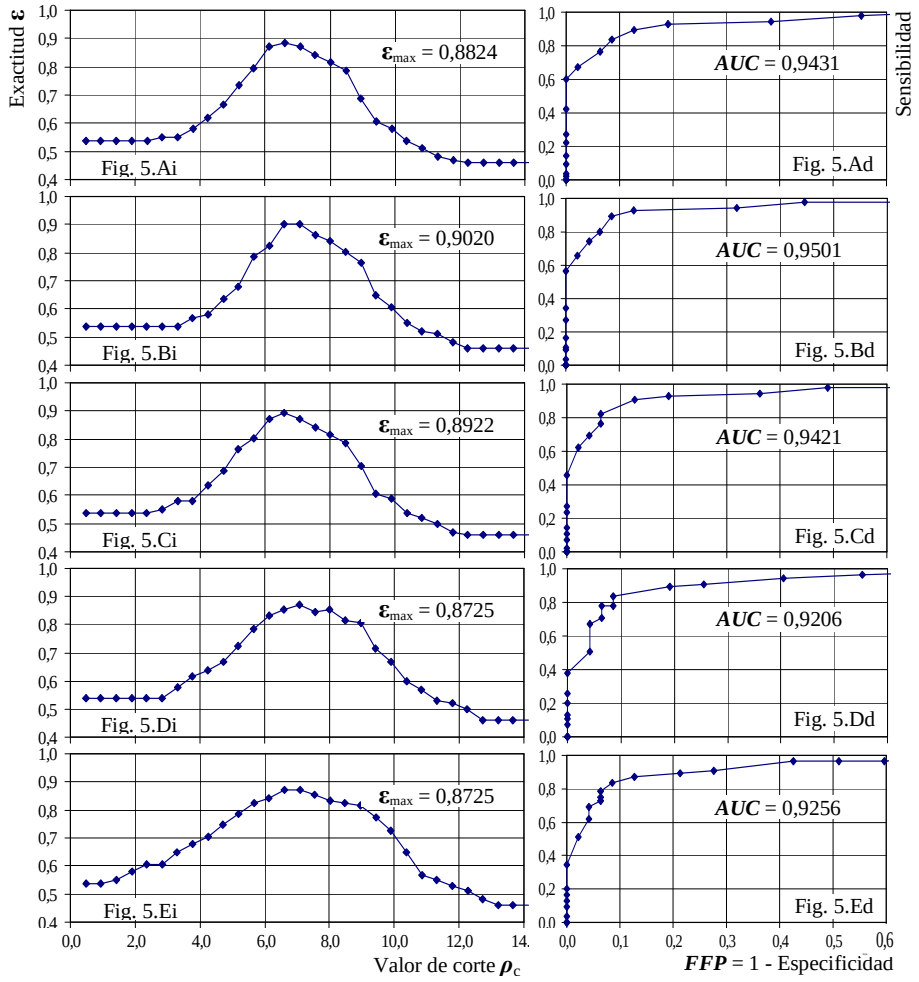
$$\boldsymbol{Exactitud = \frac{VP + VN}{VP + FN + VN + FP}} \quad (6)$$

Las curvas ROC representan la sensibilidad (Ec. 3) en función de ***FFP*** (Ec.5) y el mejor modelo es el que encierra la mayor área bajo esta curva (***AUC***) [9][10][11]. Cabe aquí destacar que las curvas ROC presentan la ventajosa propiedad de que son insensibles a los cambios en la proporción de instancias positivas y negativas que pueda haber en un lote de datos. Por su parte, el área bajo la curva ROC (***AUC***) exhibe también importantes propiedades estadísticas.

3 Presentación y discusión de los resultados obtenidos

La comparación entre el modelo básico y sus variantes se estableció en base a las áreas bajo las Curvas ROC y también se consideró la exactitud (Ec. 6) de cada modelo. El "caso de estudio" utilizado [7] consiste en un lote de 102 muestras, que incluyen 55 proyectos trazados exitosamente y 47 proyectos con resultados negativos. Estos datos ya fueron representados en los gráficos de las Figuras 2.a y 2.b. En la

Figura 5 se presentan las curvas de exactitud y curvas ROC correspondientes al modelo básico y sus variantes.



Figuras 5: Izquierda, para el modelo básico y sus variantes, representación de la exactitud **ε** en función del valor de corte ρ_c (5.Ai, 5.Bi, 5.Ci, 5.Di, 5.Ei) y derecha, representación parcial de las correspondientes Curvas ROC (5.Ad, 5.Bd, 5.Cd, 5.Dd, 5.Ed) para los mismos modelos.

Finalmente, se adopta como indicador de comparación el coeficiente de Gini (***G***), que es una medida de dispersión estadística propuesta por el italiano Corrado Gini y directamente relacionado al área bajo la curva ROC (***AUC***):

$$G = 2.AUC - 1 \qquad (7)$$

El coeficiente ***G*** se aplica al estudio de desigualdades en diversos campos, tales como la sociología, economía e ingeniería, entre otros, donde los mayores valores de ***G*** están asociados mejores desempeños, siendo su máximo posible 1 ($0 \leq G \leq 1$).

Para facilitar la comparación de los modelos se presenta en la Tabla 7 un resumen de

lo resultados obtenidos: Área bajo las curvas ROC (***AUC***), coeficiente de Gini ***G***, máxima exactitud $\boldsymbol{\varepsilon}_{max}$ y valor de corte $\boldsymbol{\rho}_c$ correspondiente a cada caso. Los valores de la segunda y cuarta columna (***AUC*** y $\boldsymbol{\varepsilon}_{max}$) son también mostrados sobre los gráficos de la Figura 5.

Tabla 7: Resumen del desempeño de los modelos de clasificación

Modelo	***AUC***	***G***	$\boldsymbol{\varepsilon}_{max}$	$\boldsymbol{\rho}_c$
A	0,9431	0,8862	0,8824	6,5997
B	0,9501	0,9002	0,9020	6,5997 - 7,0711
C	0,9421	0,8842	0,8922	6,5997
D	0,9206	0,8412	0,8725	7,0711
E	0,9256	0,8512	0,8725	6,5997 - 7,0711

Un análisis de los resultados obtenidos con los diferentes modelos sobre el caso de estudio utilizado permite hacer las siguientes consideraciones:

a) En el ***AUC*** la mayor diferencia está entre las variantes B y D, que es del 3,1%. En la exactitud la mayor diferencia del 3,3% se presenta comparando la variante B con las D y E. Esto significa que las importantes diferencias en las métricas propuestas tienen un impacto mucho menos significativo en el desempeño de los modelos.

b) El mejor valor de corte $\boldsymbol{\rho}_c$ se encuentra en todos los casos entre 6,5997 y 7,0711, lo que representa un entorno del 6,7 % tomando como referencia el mayor valor. Si se tiene en cuenta que $|\boldsymbol{\rho}_c|_{max} = 14$, se concluye que los diferentes modelos definen al valor de corte con una dispersión del 3,4 %. Aquí también las variantes en las métricas exhibieron escaso impacto en la predicción del mejor valor de corte.

c) Tanto la curva de exactitud como la curva ROC del Modelo A presentan las evoluciones más suaves, sin discontinuidades apreciables en sus pendientes.

d) Por el contrario, las variantes B a E propuestas conducen a curvas ROC con cambios súbitos de pendientes, lo que es habitual en curvas ROC de poblaciones relativamente poco numerosas como la considerada.

e) Los Modelo B y E presenta dos valores de corte que brindan la misma exactitud, el primero con la exactitud más alta y el segundo con una de las dos más bajas.

f) Los modelos que no incluyen la función sigmoidal (A, B y C) son los que presentan las mayores áreas ***AUC*** con respecto a las variantes D y E que tiene sus variables parcialmente o totalmente afectadas de la función sigmoidal.

g) Las variantes B y C, que tienen métricas lineales para las ocho variables, son las que presentan la mayor exactitud. La variante B lo hace en un intervalo de $\boldsymbol{\rho}_c$ más amplio.

h) Como se comprueba, el efecto de la función sigmoidal de tender a polarizar los valores de los argumentos no tuvo en este tipo de modelo el efecto esperado, ya que se anticipaba que contribuiría a discriminar con más facilidad los dos tipos de poblaciones de proyectos (trazables y no trazables), cosa que no evidencian los resultados.

i) Tampoco tuvieron un efecto destacado las métricas polinomiales del Modelo A, que despertaron expectativa al momento de su elección al desarrollarse este primer modelo [6], pero fueron superadas por las métricas lineales.

j) Por su relación directa con ***AUC***, el coeficiente de Gini ***G*** no aporta información nueva pero representa un indicador tradicional de la eficacia de clasificadores.

5 Conclusiones y trabajo futuro

Al procurar entender las causas que obstaculizan la aplicación efectiva de la trazabilidad en la industria del software se ingresó en un mundo complejo y apasionante, que en cierta medida responde a la estrictez de las matemáticas y al mismo tiempo es impactado por las conductas, muchas veces ambiguas, de los seres humanos. En ese contexto se identificaron ciertos factores que se consideraron determinantes, se propusieron métricas para cuantificarlos y se vienen desarrollando modelos para procurar reproducir los escenarios en los que se aplican los sistemas de trazabilidad y sus resultados. Tan pronto se pusieron a prueba los primeros modelos surgió el interrogante sobre las métricas más apropiadas para asegurar la eficacia de estas herramientas en la predicción de la trazabilidad de los proyectos de software. En este trabajo se propusieron cinco juegos de métricas y se compararon sus desempeños con un caso de estudio adoptado como referencia, llegándose a la conclusión que las métricas lineales son las más convenientes. Los próximos pasos estarán orientados a corroborar estas conclusiones con otros casos de estudio, para lo cual se trabaja paralelamente en obtener más datos de casos reales en la cantidad, calidad y variedad necesaria. La posibilidad de disponer de un modelo efectivo de predicción de trazabilidad de proyectos justifica con creces el esfuerzo que se viene realizando.

Referencias

1. Blaauboer, F., Sikkel, K., Aydin, M.: Deciding to adopt requirements traceability in practice. Proc.of the 19th Int. Conf. on Advanced Infor. Systems Engineering. Springer-Verlag (2007).
2. Kannenberg, A., Saiedian, H.: Why Software Requirements Traceability Remains a Challenge. CrossTalk: The Journal of Defense Software Engineering. July/August, 14-19 (2009).
3. Ramesh, B.: Factors influencing Requirements Traceability Practice. Communications of the ACM. 41(12), 37–44. (1998)
4. Giró, J., Vazquez, J., Meloni, B., Constable, L., Jornet, A.: Aseguramiento de la Trazabilidad en Proyectos de Desarrollo de Sistemas de Software. Proyecto de Investigación, Secretaría de Ciencia y Tecnología, Código SCyT 1214/10. (2010)
5. Giró, J., Vazquez, J., Meloni, B., Constable, L., Jornet, A.: Modelos para anticipar la factibilidad de que un proyecto de desarrollo de software sea trazable. Workshop de Ingeniería de Software, CACIC 2011. Universidad Nacional de La Plata, 837-846 (2011).
6. Giró, J., Vazquez, J., Meloni, B., Constable, L., Jornet, A.: Hacia una respuesta al interrogante de si será factible trazar un cierto proyecto de desarrollo de software. Informe Técnico 2012/01, Proyecto 1214, SCyT, FRC, UTN (2012).
7. Giró, J., Vazquez, J., Meloni, B., Constable, L., Jornet, A.: Uso del Análisis ROC para anticipar la conveniencia de trazar proyectos de software. Workshop de Ingeniería de Software, CACIC 2012. Universidad Nacional del Sur, Ciudad de Bahía Blanca (2012).
8. Powers, D.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007).
9. Fawcett, T.: An introduction to ROC analysis. Elsevier ScienceDirect, Pattern Recognition Letters, 27 - 861–874 (2006).
10. Shin Y., Huffman Hayes J., Cleland-Huang J.: A framework for evaluating traceability benchmark metrics. TR:12–001, DePaul University, School of Computing (2012).
11. Biggerstaff, B.: Comparing diagnostic tests: a simple graphic using likelihood ratios. Statistics in Medicine 19 (5) 649–663 (2000).

Análisis de rendimiento del algoritmo SGP4/SDP4 para predicción de posición orbital de satélites artificiales utilizando contadores de hardware

Federico J. Díaz[1], Fernando G. Tinetti[2,3], Nicanor B. Casas[1], Graciela E. De Luca[1], Sergio M. Martin[1], Daniel A. Giulianelli[1]

[1]Universidad Nacional de La Matanza
Florencio Varela 1903 - San Justo, Argentina
[2]III-LIDI, Facultad de Informática, UNLP
Calle 50 y 120, 1900, La Plata, Argentina
[3]Comisión de Inv. Científicas de la Prov. de Bs. As.
fedediazceo@gmail.com, fernando@info.unlp.edu.ar, {smartin, ncasas, gdeluca, dgiulian}@ing.unlam.edu.ar

Abstract. Durante los últimos 25 años, la predicción de posición orbital de los cuerpos en orbitas cercanas y medianas a la tierra, se calculo mediante el conjunto de algoritmos de la familia SGP (acrónimo de "Simplified General Perturbations"). En la última década, se produjo un incremento considerable en la cantidad de satélites artificiales, aumentando también el número de objetos inutilizados en órbita (comúnmente llamados "Basura Espacial"). Este incremento requiere un mayor esfuerzo computacional de los algoritmos utilizados. Para aprovechar en forma más eficiente los recursos computacionales actuales, puede ser necesario optimizar los algoritmos mencionados, e incluso plantear una solución paralela. El análisis aquí propuesto pretende determinar el rendimiento del algoritmo, identificando las zonas de cálculo intensivo, utilizando contadores de hardware para medir transferencias de memoria cache y rendimiento.

Keywords: SGP4, SDP4, Satélites, Contadores de hardware, Cálculo de Rendimiento, Basura espacial, Optimización, Modelado de orbitas.

1 Introducción

En forma casi ininterrumpida, durante las últimas 5 décadas, la población orbital de satélites aumento en forma gradual y constante. La red de seguimiento de satélites del gobierno de Estados Unidos, ha realizado en total desde el año 1957, el seguimiento de más de 26000 cuerpos creados por el hombre en órbita. En la figura 1.1 se ve un modelado de la cantidad de cuerpos estimados orbitando la tierra.

Esta red actualmente realiza seguimiento de 8000 cuerpos en forma simultánea (el resto ya no se encuentra en órbita), y solo el 7% de esos cuerpos son satélites

funcionales, es decir aproximadamente 560[1]. De los 8000 cuerpos entonces, 7460 aproximadamente, son lo denominado Basura Espacial ("Space Debris"). Estos cuerpos, potencialmente peligrosos, pueden provocar colisiones entre satélites funcionales, y/o reingresar a la atmosfera y caer en zonas pobladas, lo cual provocaría daños importantes en las mismas. El seguimiento constante de estos cuerpos, es primordial, además, para la actividad espacial, dado que es necesario a veces realizar correcciones de trayectorias para evitar colisiones.

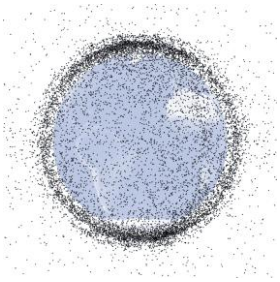


Fig. 1.1. Modelo de basura espacial y satélites en órbita cercana a la tierra. Adaptado de [1]

1.1 El problema computacional

En el año 1980 se propuso en [2] un modelo determinístico, que no requiera de un esfuerzo computacional demasiado intenso, permitiendo utilizarlo en múltiples cuerpos simultáneamente. La implementación original de este modelo, fue el algoritmo SGP[2].

El modelo actual utilizado es el SGP4, un derivado del SGP para orbitas cercanas a la tierra, y el SDP4, un derivado del SGP para orbitas superiores a los 5000 km de altura. Estos algoritmos, se utilizan para modelar orbitas cercanas a la tierra. Poseen una precisión de aproximadamente 1km, pero el error del cálculo aumenta entre 1km y 3km por cada día de predicción [3], dependiendo del cuerpo que se observe. Para reducir el error del algoritmo, se deben tomar muestras iniciales de elementos orbitales, y estas muestras se ingresan al algoritmo en un formato estándar de dos líneas[3], como veremos más adelante.

La precisión del algoritmo aumenta a medida que disminuyen los incrementos de tiempo con los que se realiza el cálculo. Pero esto conlleva a un aumento de la potencia computacional necesaria para obtener el resultado.

Mediante herramientas de performance, y contadores de hardware, se realizó un estudio sobre una implementación específica de los algoritmos, para determinar qué tipo de optimizaciones podrían aplicarse, y evaluar el rendimiento de las funciones que ejecuta.

---

[1] El número exacto no es de libre conocimiento, dado que los satélites militares no se encuentran listados en informes públicos

[2] SGP: Del inglés "Simple General Perturbations"

[3] TLE: Del ingles "Two line element"

2 El modelo de seguimiento de satélites SGP

En el presente análisis, no se pretende realizar una exposición completa del modelo, que ya está perfectamente definido en [4], pero si vamos a realizar una introducción teórica a los elementos que lo componen. También vamos a exponer las razones de porque es necesario un modelo determinístico reducido en lugar de un modelo matemático incremental que contemple todas las variables posibles.

2.1 La necesidad de un modelo simple

Si queremos realizar el seguimiento de un satélite en particular, necesitamos modelar su comportamiento orbital. En una primera aproximación, se puede realizar el cálculo de la suma de todos los vectores fuerza que se aplican sobre el satélite, y calcular instante a instante, su posición en la órbita terrestre.

Esta aproximación es lo que se conoce como integración numérica. Las fuerzas que interactúan en un satélite debajo de los 5000 km de altura (llamadas orbitas de 225 minutos), no son solo gravitatorias, también tenemos el frenado atmosférico y la presión solar, definidos como:

Frenado atmosférico. Fuerza de rozamiento que surge de la interacción del cuerpo del satélite con la atmosfera terrestre. Es significante para cuerpos en órbitas con periodos menores a 225 minutos (distancias a la superficie menores a 5000 km). [5]

Presión solar. La radiación solar genera un efecto de "presión" acumulativa en los objetos, definida por la capacidad reflectiva de la superficie donde la radiación impacta [6].

Las *ventajas* de esta solución, es que si logramos realizar incrementos pequeños de tiempo, la precisión del resultado obtenido será muy buena.

Las *desventajas* son que tenemos que conocer exactamente todas las fuerzas aplicadas al satélite en cada instante. Necesitamos muchos recursos computacionales para resolver el problema, porque si queremos calcular la posición de un satélite, dado su estado inicial, tendremos que calcular todos sus estados intermedios desde ese estado inicial.

2.2 La solución: Modelo de perturbaciones simples

En los años 80, finalmente se opto por definir un modelo que tenga en cuenta las perturbaciones en la órbita. Realizando ciertas simplificaciones, se favorece el tiempo de cálculo vs. la precisión de los resultados.

Las *perturbaciones* en una órbita, son todos aquellos efectos físicos que modifican la posición orbital de un satélite. El frenado atmosférico y la presión solar anteriormente definidos, son *perturbaciones.* Este modelo, es un modelo determinístico, que permite calcular la posición y la velocidad de un satélite para un tiempo elegido sin la necesidad de una integración numérica.

Las *consideraciones* del modelo son las siguientes: La *masa* del satélite respecto a la de la tierra es *despreciable***,** y la *órbita* del satélite se considera *cercana* a la tierra y de baja excentricidad[4]

Luego se extendió la teoría para soportar orbitas con periodos mayores a los 225 minutos. En estas orbitas, el frenado atmosférico es inexistente, pero comienzan a predominar perturbaciones como las producidas por la resonancia gravitatoria entre la el sistema tierra-luna, el sistema tierra-sol, y la presión solar. A esta modificación, que pretende abarcar todos los cuerpos artificiales orbitando la tierra, se la denominó *SGPD4*[5]

2.3 Elementos de SGPD4

En la figura 2.1, pueden apreciarse los elementos del modelo, llamados "Elementos Keplerianos" [7] [6]. Básicamente, con 6 elementos, se puede definir la posición de un objeto en una órbita terrestre para un tiempo determinado.

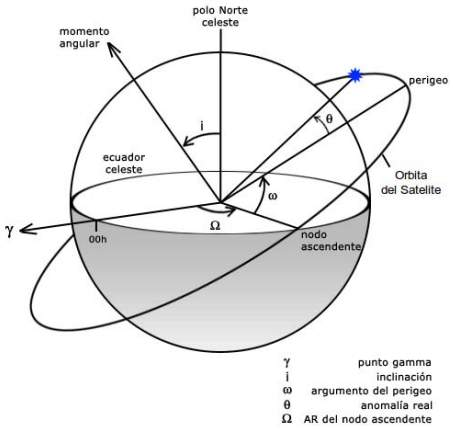


Fig. 2.1 Elementos keplerianos para un satélite en órbita. Adaptado de [8]

Donde:

Excentricidad (e) define la forma de la elipse, describiendo que tan elongada es comparada a una circunferencia (no indicada en el diagrama).

Movimiento medio es el equivalente a una velocidad media del satélite. La unidad es revoluciones orbitales por día. Está relacionada con el eje semi-mayor o el radio de la órbita. (No indicada en el diagrama).

Inclinación (i) Angulo comprendido entre el plano de la órbita, y el plano de referencia, medido en el nodo ascendente

AR[7] de nodo ascendente (Ω) orienta en forma horizontal el nodo ascendente de la elipse (donde el cuerpo realiza un tránsito ascendente respecto al plano de referencia), toando como base el punto vernal del marco de referencia.

---

[4] En oposición a una órbita de decaimiento rápido [9]

[5] Se denomina también SDP4 al algoritmo que realiza seguimiento de satélites en orbitas superiores a los 225 minutos, el nombre SGPD4 es una conjunción de ambos modelos

[6] En la referencia indicada, se consideran 8 elementos, dado que el tiempo (Época) y el frenado atmosférico(se indica como opcional), se consideran elementos del modelo

Argumento del perigeo (ω) define la orientación de la elipse en el plano orbital como un ángulo medido desde el nodo ascendente hacia el perigeo (el punto más cercano del satélite a la tierra)
Anomalía real (θ) representa el ángulo real en el plano de la elipse entre el perigeo y la posición del satélite en un tiempo determinado. A este tiempo se le llama "Época"[8]
En lugar de la anomalía real, se suele medir la anomalía media
Anomalía media es un ángulo que matemáticamente varia en forma lineal respecto del tiempo, pero no se corresponde correctamente con la anomalía real.
La *anomalía media* se puede convertir a una *anomalía real*

2.4 Limitaciones del modelo SGPD4

Como mencionamos previamente, al realizar simplificaciones en la predicción de las orbitas para optimizar el uso de recursos computacionales, estamos también, introduciendo un error en los resultados. Con tal de corregir ese error, el modelo necesita puntos de referencia para cada objeto que predice. Mediante la utilización de puntos de referencia, se pueden seguir utilizando las simplificaciones establecidas. Estos puntos de referencia, se generan mediante observaciones (directas o por radar), y se computan en un formato estándar de dos líneas, llamado TLE.

2.5 TLE: Elementos de dos líneas

Los TLE (del inglés "*Two line elements*"), se generan para cada elemento que se quiera predecir su trayectoria con el modelo SGPD4. Un TLE generado para este modelo, se realiza utilizando las mismas simplificaciones que el modelo establece, con lo cual no puede ser usado en otro modelo de seguimiento. Un TLE *solo* sirve para el modelo SGPD4. En la figura 2.2 podemos ver un ejemplo de TLE con sus elementos

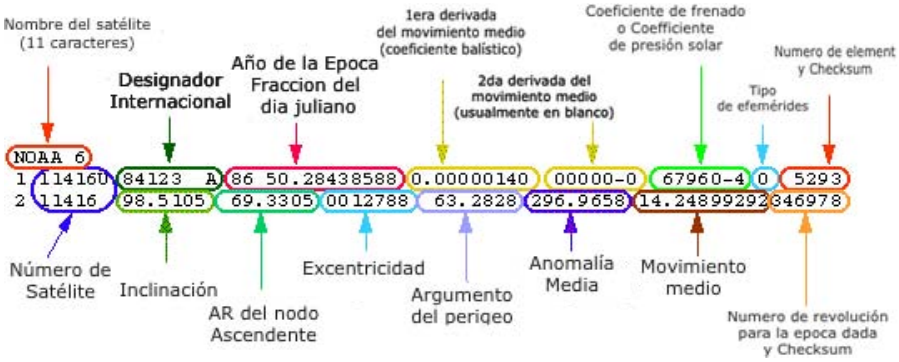


Fig. 2.2. Ejemplo de un TLE con sus elementos. Adaptado de [10]

[7] **AR:** Ascensión Recta
[8] Representa el tiempo en el cual se quiere realizar la predicción de la órbita

Estos parámetros condicen con los parámetros antes mencionados que el algoritmo SGPD4 necesita para predecir una órbita. Mediante observaciones se determinan, y se colocan en este formato. Para cada satélite que se quiera predecir su órbita, se necesita tener su respectivo TLE. [11]

La importancia del modelo, no radica tanto en su precisión, sino en el volumen de datos observacionales que se generan para cada satélite. La red de observación y seguimiento de satélites del gobierno de Estados Unidos mantiene una constante renovación de los TLE para todos los objetos mayores a 1cm en órbita terrestre. La cantidad de datos disponibles es lo que le da relevancia al modelo, con lo cual una optimización paralela podría producir mejoras significativas en el aprovechamiento de los recursos utilizados.

3 Análisis de rendimiento

Para la ejecución del modelo, se utilizo la implementación del algoritmo publicada en [12], llamada "Dundee SGPD4", y está basada en [13]. Esta implementación permite la lectura de los formatos TLE, y está escrita en el lenguaje de programación C

Para el problema en particular se decidió utilizar la información (TLE) de 12 objetos (satélites y basura espacial), obtenidos de [14]. Utilizando las simplificaciones del modelo, resulta menos costoso realizar modelados de orbitas, con lo cual podemos realizar predicciones con incrementos de tiempo pequeños, para reconstruir una órbita. Para estos doce cuerpos, se realizo una modelado de la órbita en incrementos de 0.1 segundos (dt), utilizando 144000 incrementos (n), lo que da un total de

$$dt * n = 14400\ segundos\ =\ 10\ días\ de\ predicciones \quad (1)$$

3.1 Análisis previo de la ejecución del algoritmo

En la tabla 3.1 claramente se ve, que el mayor porcentaje de tiempo de ejecución se encuentra repartido entre 3 funciones, siendo *sgdp4* la que concentra casi el 73% del tiempo de ejecución.

Tabla 1.1. Salida del comando gprof para el algoritmo Dundee SGPD4 prediciendo posiciones para 12 cuerpos (se omiten las funciones no relevantes en %)

El intervalo de muestreo es de 0.01 segundos.				
% Tiempo	Segundos acumulados	Segundos por función	llamadas	Nombre de la función
72.75	0.40	0.40	2005693	sgdp4
12.73	0.47	0.07	2005693	kep2xyz
9.09	0.52	0.05	1201421	compute_LunarSolar
5.46	0.55	0.03	1201421	SGDP4_dpper

También vemos, que la función *kep2xyz* fue llamada la misma cantidad de veces que *sgdp4*. Esta función lo que realiza es la conversión final de coordenadas de elementos keplerianos, a coordenadas cartesianas, pero no es relevante desde el punto de vista funcional del algoritmo, porque es un paso de conversión.

Otra función a tener en cuenta es *compute_LunarSolar,* ya que esta función, encargada de calcular las perturbaciones lunares y solares para satélites en orbitas mayores a los 5000 km, es llamada para 7 de los cuerpos elegidos, que se encuentran en orbitas superiores a la altitud mencionada. En la tabla 3.2 podemos ver las funciones seleccionadas para analizar

Tabla 3.2. Funciones candidatas a optimizar

% Tiempo	Segundos acumulados	Segundos por función	Llamadas	Nombre de la función
72.75	0.40	0.40	2005693	sgdp4
9.09	0.52	0.05	1201421	compute_LunarSolar

3.2 Análisis mediante contadores de hardware

Para analizar las funciones en detalle, vamos a hacer uso de la biblioteca PAPI (del inglés *"Performance Application Programmer Interface"*) [15].

Vamos a analizar ambas funciones, desde el punto de vista de la utilización de cache y cantidad de Flop de ejecución por segundo (FLOPs). De esta manera vamos a poder determinar con un número real, cual es la cantidad de operaciones estimadas necesarias para nuestra muestra de satélites. También, analizando los fallos de cache, vamos a poder utilizar esta información para detectar posibles optimizaciones a realizar.

La biblioteca PAPI, nos ofrece una cantidad de "Eventos" hardware, que podemos seleccionar para medir en áreas específicas de nuestro código. Estos eventos nos proporcionan información detallada de lo que está ocurriendo en el hardware cuando nuestro algoritmo se ejecuta. El análisis propuesto, es comparar las operaciones de punto flotante, con los fallos de cache, para comprobar si se obtiene una correlación entre estas dos medidas. Si esta correlación existe, entonces se podrá comprobar que el algoritmo necesita una optimización en memoria cache.

El análisis de MFlop que ejecuta esta función, se realiza mediante el evento *PAPI_FP_OPS*, que mide la cantidad de operaciones de punto flotante, y la función de biblioteca *PAPI_get_real_usec*, que funciona como un reloj real de máquina. Luego realizando una operación matemática, obtenemos los Flops

$$\frac{\text{Cantidad de FLOP}}{\text{Tiempo de ejecucion (s)}} = cantidad\ de\ flop\ por\ segundo(FLOPs) \quad (\mathbf{2})$$

Para el análisis de memoria, se utilizaran dos eventos, *PAPI_L1_DCM*, y *PAPI_L1_ICM*. El primero nos calcula la cantidad de fallos de cache de datos, y el

segundo la cantidad de fallos de cache en instrucciones. El evento *PAPI_L1_TCM,* es un evento que combina ambos fallos de cache en uno solo.

3.3 Análisis de las operaciones de punto flotante vs. fallos de cache de la función SDGP4

En la tabla 3.3 podemos ver, que en el sexto cuerpo (SGP4 simple, marcado en el recuadro), se observa un comportamiento significativo. Se producen pocos fallos de cache, pero se obtiene un alto nivel de MFlops. Esto ocurre porque la simplicidad del modelo permite una alta reutilización de las variables intervinientes, así como también un bajo tiempo de ejecución, maximizando la eficiencia del algoritmo.

Tabla 3.3. Tiempo (en microsegundos), FLOP, Flops calculados (expresados en MFlops) y fallos de cache L1 para la función SGDP4.

Modelo	**SGP4 normal**	**SGP4 normal**	**SGP4 normal**	**SGP4 normal**
Tiempo (µs)	413673	408677	409513	407684
FLOP	45718720	45260232	45426294	45508607
MFlops	111	111	111	112
Fallos de cache L1	4908969	4415024	4502586	4457233
Modelo	**SDP4 resonante**	**SGP4 simple**	**SDP4 resonante**	**SDP4 normal**
Tiempo (µs)	491014	76416	487726	521072
FLOP	69287673	5443941	69478753	76542298
MFlops	141	71	142	147
Fallos de cache L1	8279346	396473	7853828	9890781
Modelo	**SDP4 normal**	**SDP4 normal**	**SDP4 sincrónico**	**SDP4 sincrónico**
Tiempo (µs)	481717	522181	508021	499020
FLOP	67755484	79058341	77300346	75665619
MFlops	141	151	152	152
Fallos de cache L1	7460974	10583237	10314707	10449186

De aquí podemos concluir, que a mayores perturbaciones, mayores fallos de cache se obtendrán de la implementación. El cuerpo 5, y los cuerpos del 7 al 12, responden a orbitas superiores a los 225 minutos. En estas orbitas, tenemos un problema pseudo-incremental, en el cual tenemos que añadir las perturbaciones requeridas. El indicio del comportamiento del sexto cuerpo, nos marca el camino de donde el modelo necesita ser optimizado. Esto es, en el cálculo de las perturbaciones.

3.4 Análisis de la Función *compute_LunarSolar*

Para analizar esta función vamos a realizar una pequeña diferencia, y vamos a tener en cuenta por separado los fallos de cache de código, y de cache de datos. En la tabla 3.4 podemos ver que la diferencia de MFlops calculados entre el modelo SGP4 y el modelo SDP4, es básicamente el cálculo de las perturbaciones.

Tabla 3.4. Tiempo (en microsegundos), FLOP, MFlops calculados y fallos de cache L1 de código y datos para la función *compute_LunarSolar*. Se omiten los primeros 4 cuerpos, ya que responden al modelo SGP4 y no utilizan estas perturbaciones.

Modelo	**SDP4 resonante**	**SGP4 Simple**	**SDP4 resonante**	**SDP4 normal**
Tiempo (µs)	364346	0	364492	365418
FLOP	19758109	0	19780697	19742655
MFlops	54	0	54	54
Fallos de cache L1 código	2775539	0	2720791	2640799
Fallos de cache L1 datos	237326	0	186601	233858
Modelo	**SDP4 normal**	**SDP4 Normal**	**SDP4 sincrónico**	**SDP4 sincrónico**
Tiempo (µs)	365745	363005	364334	361646
FLOP	21028810	19644869	19787661	19264844
MFlops	57	54	54	53
Fallos de cache L1 código	2189155	2868253	2693908	2776276
Fallos de cache L1 datos	152355	250608	260512	273745

Este incremento de MFlops es del orden del 50% aproximadamente, debido a que los MFlops de esta función, deben sumarse a los MFlops de la función SDGP4.Los fallos de cache de código, pueden mejorarse mediante la agrupación de las soluciones.

Si se hace un análisis del TLE previo a la ejecución del algoritmo, se puede determinar qué tipo de modelo orbital utiliza. Sabiendo esto, podemos reordenar la ejecución de la predicción, para aprovechar el código de ejecución que se encuentra en cache, y solo los datos intervinientes se modificarían.

4 Conclusiones y trabajo futuro

Realizar optimizaciones de modelos orbitales, es una tarea compleja. Generalmente estos modelos involucran operaciones matemáticas complejas, y requieren un esfuerzo computacional significativo. En este primer análisis de rendimiento, se intento lograr obtener el punto de base para realizar optimizaciones futuras. Ese punto, debería ser la optimización de las funciones que calculan las perturbaciones

solares y lunares. Específicamente, la función *compute_LunarSolar* es la candidata principal a ser optimizada.

Inicialmente el modelo original, intentó ser una forma determinística de fijar posiciones de objetos en orbitas cercanas. Con el paso del tiempo, fue necesario introducir modificaciones de cálculo incremental para orbitas profundas, que trajeron consigo un incremento en la potencia computacional necesaria para resolver el problema.

Con este trabajo se intenta dar un paso hacia adelante en la dirección de la optimización del modelo predictivo. Al realizar un análisis objetivo desde el punto de vista de las operaciones en punto flotante, y del manejo de cache, la función mencionada produjo resultados que podrían indicar que es posible obtener mejoras en una optimización futura de la misma.

Se propone que el mayor esfuerzo de análisis, debería concentrarse en el desafío de obtener versiones paralelas de todas las perturbaciones calculadas posibles. Inclusive, la posibilidad de distribuir el trabajo computacional utilizando GP-GPU para este fin.

Referencias

1. Earth Observatory, NASA: Space Debris, http://earthobservatory.nasa.gov
2. Felix R. Hoots, Ronald L. Roehrich: SPACETRACK REPORT NO.3: Models for Propagation of NORAD Element Sets (Diciembre 1980)
3. T.S. Kelso: Real-World Benchmarking. Satellite Times, 3, no. 2. pp. 80-82 (Noviembre/Diciembre 1996)
4. Felix R. Hoots, Ronald L. Roehrich: SPACETRACK REPORT NO.3: Models for Propagation of NORAD Element Sets (Diciembre 1980)
5. E.M Gaposchkin and AJ. Coster: Analysis of Satellite drag. Lincoln Laboratory Journal, Massachusetts Institute of technology, vol. 1, no 2. (1998)
6. Nichols, E.F & Hull, G.F. (1903) The Pressure due to Radiation, The Astrophysical Journal,Vol.17 No.5, p.315-351.
7. The Radio Amateur Satellite Corporation, http://www.amsat.org/amsat/keps/kepmodel.html
8. Human Space Flight, NASA: Definition of Two-line Element Set Coordinate System, http://spaceflight.nasa.gov
9. John Kennewell: Satellite Orbital Decay Calculations. IPS Radio and space services, Australian Government, Bureau of Meteorology (1999)
10. Instrument Working Group, NASA: Keplerian Elements, http://earth-www.larc.nasa.gov/ceresweb/IWG/intro.html
11. T.S. Kelso: Frequently Asked Questions: Two-Line Element Set Format. Satellite Times, vol. 4, no. 3, pp 52-54 (Enero 1998).
12. Dundee satellite receiving station, Dundee University, http://www.sat.dundee.ac.uk/
13. Felix R. Hoots, Ronald L. Roehrich: SPACETRACK REPORT NO.3: Models for Propagation of NORAD Element Sets (Diciembre 1980)
14. Space-Track organization, United States Government, https://www.space-track.org/
15. Performance Application Programmer Interface, http://icl.cs.utk.edu/papi/
16. T.S. Kelso: Orbital Propagation, Part I. .Satellite Times, vol. 1, no. 1, pp 70-71 (Septiembre/Octubre 1994)
17 H. Karttunen, P. Kröger, H. Oja, M. Poutanen, K. J. Donner: Fundamental Astronomy, Springer - Verlag, Berlín (2007)
18 Fernando G. Tinetti, Armando De Giusti: "Procesamiento Paralelo. Conceptos de Arquitecturas y Algoritmos", Editorial Exacta, (1998).

X WORKSHOP BASES DE DATOS Y MINERÍA DE DATOS

- WBDDM -

CACIC2013

X WORKSHOP BASES DE DATOS Y MINERÍA DE DATOS

- WBDDM -

ID	Trabajo	Autores
5612	**A Novel, Language - Independent Keyword Extraction Method**	Waldo Hasperué (UNLP), César Estrebou (UNLP), Laura Lanzarini (UNLP), Germán Aquino (UNLP)
5618	**Modelo de Procesos para la Gestión de Requerimientos en Proyectos de Explotación de Información**	Maria Florencia Pollo Cattaneo (UTN-FRBA), D. Mansilla (UTN-FRBA), C. Vegega (UTN-FRBA), Patricia Pesado (UNLP), Ramón Garcia Martinez (UNLA), Paola Britos (UNRN)
5626	**Efecto de los trending topics en el Volumen de Consultas a los Motores de Búsqueda**	Santiago Ricci (UNLu), Gabriel Tolosa (UNLu)
5721	**Propuesta de Métricas para Proyectos de Explotación de Información**	Diego Basso, Darío Rodriguez (UNLA), Ramon Garcia Martinez (UNLA)
5771	**Fractalizing Social Networks**	Silvia Cobialca (OTRA), Juan María Ale (UBA)
5819	**Determinación de género y edad en blogs en español mediante enfoques basados en perfil**	Darío Funez (UNSL), Leticia Cagnina (UNSL), Marcelo Errecalde (UNSL)
5828	**Una Extensión del FHQT Temporal para Distancias Continuas**	Andrés Pascal (UTN), Anabella De Battista (UTN), Norma Edith Herrera (UNSL), Gilberto Gutierrez (UBB)
5875	**New Deletion Method for Dynamic Spatial Approximation Trees**	Fernando Kasián (UNSL), Verónica Ludueña (UNSL), Nora Reyes (PUC Rio), Patricia Roggero (UNSL)

X WORKSHOP BASES DE DATOS Y MINERÍA DE DATOS

- WBDDM -

ID	Trabajo	Autores
5825	**Prototipo de búsqueda y comparación que aplica técnicas de recuperación de información en bases de datos relacionales**	Claudio Camacho (UNSE), Walter Singer (UNSE), Rosanna N. Costaguta (UNSE)

A Novel, Language-Independent Keyword Extraction Method

Germán Aquino[1], Waldo Hasperué[1,2], César Estrebou[1] and Laura Lanzarini[1]

[1] III-LIDI. School of Computer Science. UNLP. Argentina
[2] CONICET scholarship
{gaquino, whasperue, cesarest, laural}@lidi.info.unlp.edu.ar

Abstract. Obtaining the most representative set of words in a document is a very significant task, since it allows characterizing the document and simplifies search and classification activities. This paper presents a novel method, called LIKE, that offers the ability of automatically extracting keywords from a document regardless of the language used in it. To do so, it uses a three-stage process: the first stage identifies the most representative terms, the second stage builds a numeric representation that is appropriate for those terms, and the third one uses a feed-forward neural network to obtain a predictive model. To measure the efficacy of the LIKE method, the articles published by the Workshop of Computer Science Researchers (WICC) in the last 14 years (1999-2012) were used. The results obtained show that LIKE is better than the KEA method, which is one of the most widely mentioned solutions in literature about this topic.

Keywords: Text Mining, Document characterization, Back-propagation, WICC.

1 Introduction

Text mining presents interesting challenges to solve, since the lack of structure in the texts analyzed makes it difficult to extract information from them. Nowadays, given the large number of texts that are published each day, be these scientific articles, books, journals, periodicals or web pages, facing these challenges can prove to be interesting, as well as developing strategies that allow obtaining information from relevant texts.

One way of briefly describing the topic of a document is by means of a list of keywords. The keywords in a document are of the utmost importance, since they allow carrying out several tasks, such as searching for a specific topic, classifying documents, clustering [1], summarization [2] [3] [4], etc.

Even though most of the times the author of the document is the one in charge of proposing the list of keywords, as in the case of scientific publications, there are other times when this list is not present at all and, therefore, it would be interesting to have an automated method that can propose a list of keywords by analyzing the text of the document.

Within text mining, there have been various alternatives proposed for the task of extracting keywords. There are statistical methods that typically do not have prior training with the documents; in such cases, only statistical information is collected from the words that are present in the document to identify which of them can be chosen as keywords. The most widely used statistical methods include TF-IDF [5] [6] [7] [8], word co-occurrence [9], etc.

On the other hand, there are machine learning-based methods that, from a given corpus, carry out a training process and generate a model that allows performing classifications afterwards or, in the case of keyword extraction, establishing which words in the document are candidates to be chosen as keywords. In these cases, each document in the initial corpus must have a list of keywords that are used as positive cases during training. Some of the machine learning methods used in this type of tasks are Naïve Bayes [8] [10], Support Vector Machine [11], etc.

The methods that analyze the linguistic aspects of the documents are those that offer the most interesting solutions, since they combine lexical analysis, syntactic analysis, etc. [12] [13]; however, their disadvantage is that they are strongly dependant on the language used to write the documents.

One of the main concepts pertaining to the specific task of extracting keywords from documents is that of n-grams. Any word within a document is a unigram, while any sequence of two or more words forms an n-gram, where *n* indicates the number of words in the sequence. When extracting keywords, any n-gram in the document is a potential keyword for that document. Most of the techniques that carry out this task perform calculations and measurements on each n-gram in the document, and then process them by means of a machine learning technique [14] or by assigning a given score [15] to obtain a model that can be used as predictor for future documents.

In order to extract keywords, some methods require a corpus from which to generate a first model [7] [8], while others do so from a single document [9]. There are techniques that carry out numeric and/or statistical calculations for all n-grams in the document [16] [17], while others exploit certain linguistic information [12] [13]. Most of the existing strategies pre-process the documents with stemming and word filtering techniques by means of a stop-word list.

In this paper, a novel method is proposed, called LIKE (Language Independent Keyword Extraction), which uses texts from documents from a given corpus to obtain a model that can be used to extract keywords. To this end, it uses a three-phase algorithm. The first phase consists in extracting any n-grams that are detected as candidates for keyword, the second phase calculates a set of numerical features for each n-gram that was detected, and the third and final stage uses those features to produce a model by training a feed-forward network. This trained network is used as model to decide, given a new document, possible keywords. LIKE is independent from the source language of the documents, provided that the same language is maintained throughout each individual document, since it does not carry out the usual pre-processing steps of stemming or word filtering using a stop-word list.

This article is organized as follows: in Section 2, LIKE is described; in Section 3, the results obtained in the experiments carried out are presented; and in Section 4, conclusions and future work are presented.

2 LIKE

The method proposed in this paper, called LIKE, is a three-phase method that allows extracting a list of keywords by analyzing the text in a document. LIKE is independent from the source language of the documents, since it does not carry out the pre-processing steps of stemming or word filtering using a stop-word list. LIKE analyzes the documents in a document corpus to train a back-propagation network that will then be used as model to determine the list of keywords for new documents.

LIKE starts by identifying all existing n-grams in each document in the corpus. Since the number of all possible n-grams would be excessively high, a strategy is required to reduce this number. In this proposal, the method presented by [18] is used, since it allows identifying a much lower number of n-grams.

During the second stage, each n-gram obtained in the previous stage is transformed into a features numeric vector; these vectors are labeled as one of two classes of data. One class includes the vectors corresponding to those n-grams that are part of the list of keywords proposed by the author(s) of the document, while the other class is formed by the vectors corresponding to all remaining n-grams. The third stage consists in using these two data classes to train a back-propagation network.

2.1 Phase 1. N-gram Extraction

The first phase of the method proposed consists in identifying the n-grams in the corpus. For each document, all existing n-grams are extracted. An n-gram is considered to be valid if it is formed by consecutive words within the same sentence with no punctuation marks between any of them.

In general, the number of existing n-grams is excessively high. For the tests carried out for this work, which has a corpus of 96 documents, more than 580,000 n-grams can be extracted. Therefore, a strategy to identify a lower number of n-grams is required. In this proposal, the algorithm presented in [18] is used. This strategy, inspired in the Apriori algorithm, builds sets of elements from other smaller sets. In this algorithm, the maximum n value (number of words in the n-gram) and the minimum occurrence frequency for each n-gram have to be determined. N-grams are built from the (n-1)-grams that meet the requirement of a minimum specified frequency. To do this, it is assumed that an n-gram whose frequency is k is built from the intersection of two (n-1)-grams whose frequency is at least k, i.e., an n-gram cannot be more frequent than its parts. For each n-gram, the first n-1 and last n-1 words are taken, and it is checked that these (n-1)-grams meet the minimum allowed frequency criterion. If this criterion is not met, the n-gram is discarded. Finally, there are n runs on the text, first to obtain 1-grams, then 2-grams based on these, then 3-grams, and so forth.

The use of this strategy in LIKE allows identifying a low number of n-grams in each document. In the experiments that were carried out for this work, the total number of n-grams for the entire corpus was reduced to little more than 70,000.

The result of this phase is a list of n-grams, which are then labeled. Once this list of keywords is known for each document, a label is assigned to each n-gram to indicate if the n-gram is a keyword or not. Thus, a two-class set of data is generated.

2.2 Phase 2. N-gram Characterization

The purpose of this phase is converting each of the n-grams that were identified in the previous phase into a features vector. In this article, we propose that the eight features detailed below are calculated.

i) TF (Term Frequency): TF is perhaps, together with IDF, one of the descriptors most widely-used to characterize n-grams. Term Frequency is the number of times the n-gram occurs in the document divided by the total word count of the document.

ii) IDF (Inverse Document Frequency): It is the ratio between the number of documents in the corpus that include the n-gram $d(g)$ and the total number of documents D.

$$IDF(g) = \log(\frac{D}{d(g)})$$

iii) First occurrence of the term: This represents the relative position of the first occurrence of the n-gram. It is calculated as the number of words before the first occurrence of the n-gram divided by the total word count of the document.

iv) Position within the sentence: This is the relative position of the n-gram in the sentence that contains it. The same as the previous one, it is calculated as the number of words before the occurrence of the n-gram in the sentence divided by the total word count of the sentence itself. If the same n-gram appears several times in different sentences in the same document, then all n-gram occurrences are averaged.

v) Part of the title: This feature is a binary value that indicates if the n-gram appears in the title of the document or not.

vi) Part of the n-gram present in the title: This feature (only valid for n-grams with two or more words) counts the number of words in the n-gram that also appear in the title, regardless of the order of the words in the title. This number of occurrences is normalized by the number of words in the n-gram. In the case of unigrams, the same as the previous feature, this is a binary value that indicates either presence or absence.

vii) NSL (Normalized Sentence Length): This is the length of the sentence where the n-gram appears divided by the length of the longest sentence in the document. If the n-gram appears in more than one sentence in the document, all occurrences are averaged.

viii) Z-score: This is a statistical measure that normalizes the frequency of the n-gram. It requires knowing the mean and standard deviation of the frequency for each n-gram.

$$Z\text{-}score(g) = \frac{freq(g) - \mu}{\sigma}$$

If the n-gram appears in more than one sentence in the document, all occurrences are averaged.

Of the eight features proposed, only two (IDF and Z-score) require the corpus in order to be calculated.

The result of this phase is a features vector for each of the n-grams identified in the previous phase.

2.3 Phase 3. Creating the Model

The third phase of the method proposed consists in creating a model that can learn from a given corpus and allows classifying n-grams from new documents as possible keywords or not. The prediction model is built by training a back-propagation network.

The problem that arises when trying to use the set of vectors obtained in the previous phase as data for training the back-propagation network is that the classes in this data set are not balanced, since the "not a keyword" class has a lot more elements than the "is a keyword" class. In the corpus used to carry out the experiments presented here, the ratio of elements in both classes was approximately 150 to 70,000.

When there is a data set with unbalanced classes, training a back-propagation network is not an easy task, since the training set prevents the generation of a model that can accurately predict the data in the minority class. In light of this problem, several solutions have been proposed ([19] [20] [21]). In particular, the solution described in [21] proposes that, before the training process, a clustering operation is performed on the data in the larger class in order reduce its number of elements. In this work, the idea in [21] is used – the data in the larger class are clustered using the k-means algorithm.

Be u the number of data present in the minority class, the value of k is then established as $k=u/10$. A clustering of k clusters is performed, and 10 random elements are extracted from each resulting cluster. These $10*k$ elements thus selected form a new data set that replaces the original data from the majority class. Following this methodology, the back-propagation network can be trained using a data set whose classes have similar numbers of elements.

To train the back-propagation network, the classic algorithm is used. After several tests and empirical observation, a decision was made to use seven neurons for the hidden layer, the logsig function as transfer function for the hidden layer, the tansig function for the output layer, an alfa of 0.25, and a maximum of 1,000 iterations. The best results were obtained with this configuration.

The result of this training process is a model that can predict keywords for new documents. The procedure to establish the keywords for a new document is as follows: first, the n-grams are extracted as described in Section 2.1; then, features vectors are calculated for each n-gram as explained in Section 2.2; and finally, these new vectors are presented to the trained network to determine if a given n-gram is a keyword for the document or not.

3 Results

The method proposed here was tested using as corpus all papers submitted to WICC (http://redunci.info.unlp.edu.ar/wicc.html) between 1999 and 2012. Only those articles written in English were included in the corpus (96 articles). The rationale for using only articles written in English was that, at a later stage, the results obtained with this method would be compared with those obtained with other keyword extraction method that is widely used in the literature: KEA [8]. Even though KEA

can be adapted to work with languages other than English, since it depends on a stemmer and a stop-word list, those developed by the authors were used, which are in English.

KEA [8] is an automated keyword extraction algorithm that identifies candidate words by using lexical methods to calculate a set of features, and then apply an automated learning algorithm that allows predicting which candidates are good keywords.

The same as LIKE, KEA builds a prediction model using a training corpus with specified keywords, and it then uses this model to extract keywords from new documents.

KEA allows the free extraction of keywords, as well as the extraction of keywords using a vocabulary list that is controlled by means of a thesaurus. For these tests, the first mode was used, establishing as parameter a number of three keywords per document. In order to train KEA, a stop-word list and a stemmer are required. The stop-word list contains words of low semantic content (conjunctions, articles, prepositions, etc.) that should not be considered as keyword candidates.

The first step in the KEA method consists in filtering out the words that appear in the stop-word list, and then apply a stemming process to reduce to their syntactic root all n-grams that were not filtered out. The next step is to calculate the features of all candidate words, which include: TF-IDF, the initial position of the n-gram in the text and the length of the n-gram (the number of individual words that form the n-gram). Based on this representation, KEA uses Naïve Bayes as learning algorithm.

Both LIKE and KEA were trained using the same corpus. From that corpus, some documents were selected randomly for the training stage and the rest were used for testing. For each test, accuracy, recall and f-measure are calculated.

Both methods were run with the 10-fold cross-validation procedure, and average accuracy, recall and f-measure were obtained. The 10-fold cross-validation procedure was run 30 separate times with both methods in order to measure the statistical significance of the various results obtained.

One of the greatest disadvantages of LIKE (also present in KEA) is that, for each n-gram, two features are calculated whose result depends on the entire corpus (IDF and Z-score). Depending on a corpus for calculating features is not desirable, so two versions of the LIKE method were run – LIKE-8, which uses the eight features proposed in this article (see Section 2.2), and LIKE-6 which uses only the six features that do not depend on the corpus (i.e., all but IDF and Z-score).

Table 1 shows the average accuracy, recall and f-measure for the 30 separate runs with LIKE-8, LIKE-6 and KEA. With these results, a statistical test was carried out to determine the statistical significance for LIKE-8 vs. KEA, LIKE-6 vs. KEA and LIKE-8 vs. LIKE-6 (Table 2). As it can be seen in Table 2, both LIKE-8 and LIKE-6 achieved better results than KEA, while the version that used all eight attributes improved only accuracy and f-measure results when compared to the version that used only those six that are not corpus-dependent.

Table 1. Average precision, recall and f-measure for LIKE-8, LIKE-6 and KEA (standard deviation indicated between parentheses).

	LIKE-8	LIKE-6	KEA
Precision	0.76 (0.051)	0.65 (0.101)	0.52 (0.006)
Recall	0.75 (0.094)	0.72 (0.141)	0.37 (0.004)
f-measure	0.74 (0.053)	0.68 (0.116)	0.43 (0.005)

Table 2. Results of the statistical significance for precision, recall and f-measure for LIKE-8 vs. KEA, LIKE-6 vs. KEA and LIKE-8 vs. LIKE-6. For α=0.01 the "+" sign indicates that the result is statistically significant, while the "–" sign indicates that there is no statistical significance (p-value indicated between parenthesis).

	LIKE-8 vs. KEA	LIKE-6 vs. KEA	LIKE-8 vs. LIKE-6
Precision	+ ($5.48\text{x}10^{-22}$)	+ ($5.19\text{x}10^{-08}$)	+ ($4.12\text{x}10^{-06}$)
Recall	+ ($1.50\text{x}10^{-19}$)	+ ($4.08\text{x}10^{-14}$)	- (0.4832)
f-measure	+ ($2.18\text{x}10^{-24}$)	+ ($2.51\text{x}10^{-12}$)	+ (0.0096)

4 Conclusions and Future Work

The novel automated method LIKE for extracting keywords from the text of a set of documents has been presented. This method extracts n-grams from the documents and then calculates a series of features to convert them into numeric vectors. It then uses these vectors as data to train a back-propagation network and thus obtain a model that works as predictor and that can be used to extract keywords from new documents.

In this paper, the calculation of eight features is proposed for each n-gram, with only two of these being dependent on the entire corpus. LIKE was trained using the eight features, and then a second test was carried out using only the six features that do not depend on the corpus. Articles written in English submitted to the WICC between 1999 and 2012 were used for the experiments. The results obtained were compared with KEA, and it was shown that both the six-feature and the eight-feature LIKE models were better. When comparing the results obtained with both versions of LIKE, using all eight features turned out to be superior than using just six when calculating precision and f-measure, while for the recall parameter, neither of the versions appeared to be better than the other.

The main advantage of the method presented here is that it does not depend on the language of the texts analyzed, since it does not pre-process them because it does not use stemming or a stop-word list.

As future work, it would be interesting to study in detail the n-grams that are negatively classified so as to determine their nature and analyze the possibility of detecting grammar structures that help improve the performance of the method. Also, if less candidate n-grams are identified, the majority class of negative cases would be reduced and this could possibly lead to being able to omitting the clustering stage before training the network.

Another aspect to be studied in relation to the method proposed here is the possibility of assigning keywords from a list of controlled vocabulary. Different authors may choose different key words in articles dealing with the same topic, so it would be interesting if an automated assignment method were available to assign key words from a list of controlled vocabulary. This would ensure that documents on related topics would have the same key words, which would in turn improve the results obtained in future searches, classifications or statistical analyses.

References

1. Tonella, P., Ricca, F., Pianta, E., Girardi, C.: Using keyword extraction for Web site clustering. In: Conference Using keyword extraction for Web site clustering, pp. 41 - 48. (2003)
2. D'Avanzo, E., Magnini, B., Vallin, A.: Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004. Proceedings of the 2004 Document Understanding Conference (2004)
3. Wan, X., Yang, J., Xiao, J.: Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. In: Conference Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction, pp. 552-559. (2007)
4. Zha, H.: Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In: Conference Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, pp. 113-120. (2002)
5. Islam, M.R., Islam, M.R.: An improved keyword extraction method using graph based random walk model. 11th International Conference on Computer and Information Technology 225-229 (2008)
6. Kaur, J., Gupta, V.: Effective Approaches For Extraction Of Keywords. International Journal of Computer Science Issues 7, (2010)
7. Liu, Y., Ciliax, B.J., Borges, K., Dasigi, V., Ram, A., Navathe, S., Dingledine, R.: Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering. Proc IEEE Comput Syst Bioinform Conf 394-404 (2004)
8. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: practical automatic keyphrase extraction. In: Conference KEA: practical automatic keyphrase extraction, pp. 254-255. (1999)
9. Matsuo, Y., Ishizuka, M.: Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information. In: Conference Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information, pp. 392-396. (2003)
10. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C.: Domain-specific keyphrase extraction. proc. Sixteenth International Joint Conference on Artificial Intelligence 668--673 (1999)
11. Wu, C., Marchese, M., Wang, Y., Krapivin, M., Wang, C., Li, X., Liang, Y.: Data Preprocessing in SVM-Based Keywords Extraction from Scientific Documents. Fourth International Conference on Innovative Computing, Information and Control (ICICIC), pp. 810 - 813 (2009)
12. Csomai, A., Mihalcea, R.: Linguistically Motivated Features for Enhanced Back-of-the-Book Indexing. Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue (2008)

13. Kireyev, K.: Semantic-based estimation of term informativeness. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics 530-538 (2009)
14. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. Proceedings of the 2003 conference on Empirical methods in natural language processing 216-223 (2003)
15. Wang, C., Zhang, M., Ru, L., Ma, S.: An Automatic Online News Topic Keyphrase Extraction System. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 1, 214-219 (2008)
16. HaCohen-Kerner, Y., Gross, Z., Masa, A.: Automatic extraction and learning of keyphrases from scientific articles. Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing 657-669 (2005)
17. Hu, X., Wu, B.: Automatic Keyword Extraction Using Linguistic Features. Sixth IEEE International Conference on Data Mining Workshops 19 - 23 (2006)
18. Fürnkranz, J.: A study using n-gram features for text categorization. Austrian Research Institute for Artifical Intelligence 3, 1-10 (1998)
19. Castro, C.L., Braga, A.P.: Novel Cost-Sensitive Approach to Improve the Multilayer Perceptron Performance on Imbalanced Data. IEEE Transactions on Neural Networks and Learning Systems 24, 888-899 (2013)
20. Lin, M., Tang, K., Yao, X.: Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification. IEEE Transactions on Neural Networks and Learning Systems 24, 647 - 660 (2013)
21. Zhang, Y.-p., Zhang, L.-N., Wang, Y.-C.: Cluster-based majority under-sampling approaches for class imbalance learning. 2nd IEEE International Conference on Information and Financial Engineering (ICIFE) 400-404 (2010)

Modelo de Procesos para la Gestión de Requerimientos en Proyectos de Explotación de Información

Pollo-Cattaneo, M. F.[1,2], Mansilla, D [2],Vegega, C [2], Pesado, P.[3], García-Martínez, R.[4], P. Britos, P.[5]

[1] Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. Argentina.
[2] Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional. Argentina.
[3] Instituto de Investigaciones en Informática LIDI. Facultad de Informática. Universidad Nacional de La Plata. Argentina.
[4] Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina
[5] Grupo de Investigación en Explotación de Información. Laboratorio de Informática Aplicada. Universidad Nacional de Río Negro. Argentina.
fpollo@posgrado.frba.utn.edu.ar, rgarcia@unla.edu.ar, paobritos@gmail.com

Resumen. Los Proyectos de Explotación de Información tienen por objetivo la transformación de los datos que recopila el proceso de negocio, en conocimiento útil para la toma de las decisiones. Los procesos asociados a los proyectos tradicionales de construcción de software no pueden utilizarse, debido a que estos procesos están orientados a la construcción de un producto diferente, que es el software. En este contexto, se presenta un proceso de gestión de requerimientos para Proyectos de Explotación de Información, que hace hincapié en las fases de definición de proyectos, educción del negocio, conceptualización del negocio e identificación de los procesos de explotación de información asociados al Proyecto de Explotación de Información.

Palabras Claves: Elicitación de Requerimientos. Metodología. Proceso. Explotación de Información. Ingeniería de Requerimientos.

1. Introducción

La Inteligencia de Negocios propone un abordaje interdisciplinario (dentro del que se encuentra la Informática) que se centra en generar conocimiento que contribuya con la toma de decisiones de gestión y generación de planes estratégicos en las organizaciones [1]. La Explotación de Información (EdI) es la sub-disciplina de la Informática que aporta a la Inteligencia de Negocios las herramientas de análisis y síntesis para extraer conocimiento no trivial que se encuentra (implícitamente) en los datos disponibles de diferentes fuentes de información [2]. Para un experto, o para el responsable de un Sistema de Información, normalmente no son los datos en sí lo más relevante, sino el conocimiento que se encierra en sus relaciones, fluctuaciones y dependencias.

Si bien existen metodologías que acompañan el desarrollo de Proyectos de Explotación de Información que se consideran probadas y tienen un buen nivel de madurez en cuanto al desarrollo del proyecto entre las cuales se destacan CRISP-DM [4], P3TQ [5] y SEMMA [6], estas metodologías dejan de lado aspectos operativos y de gestión de proyecto [7]. Así, por ejemplo, en la metodología CRISP-DM la primera fase busca identificar y comprender los aspectos del negocio relacionados al proyecto que se está realizando, pero no define técnicas, métodos ni herramientas para obtener esta información ni los medios necesarios para realizar su documentación.
En este contexto, este trabajo tiene como objetivo sistematizar el cuerpo de conocimientos existente en la Ingeniería en Software y la Ingeniería del Conocimiento para sentar las bases para el desarrollo de una Ingeniería de Requisitos con particular énfasis en Proyectos de Explotación de Información. De esta manera se propone un Modelo de Procesos para la Gestión de Requisitos en Proyectos de EdI. Para ello primero se describe el problema detectado (sección 2). Luego se definen las soluciones existentes junto con la nueva propuesta (sección 3), aplicando esta última en una prueba de concepto para su validación (sección 4). Finalmente se indican las conclusiones obtenidas y futuras líneas de trabajo (sección 5).

2. Definición del Problema

Aunque la Ingeniería en Software y la Ingeniería del Conocimiento [8] proveen muchos métodos, técnicas y herramientas, estos no son útiles ya que no se ocupan de los aspectos específicos que poseen los Proyectos de Explotación de Información. Mientras, que las herramientas tradicionales de elicitación de la Ingeniería en Software se enfocan en la descripción de los diferentes tipos de requerimientos haciendo hincapié en las características (funcionales o no) que debe cumplir el producto software final [9], un proyecto EdI no busca la construcción del sistema software sino la aplicación de un proceso que convierta los datos disponibles en conocimiento. Por lo tanto, los requerimientos en este tipo de proyecto, se encontrarán relacionados a identificar y describir los objetivos del proyecto junto con su relación con los objetivos del negocio de la organización donde se está realizando el proyecto. Además es necesario realizar un reconocimiento inicial de las fuentes de información disponibles en la organización identificando cuáles fuentes se encuentran informatizadas (en repositorios de datos) y cuáles no. Como resultado del análisis conjunto de los objetivos del proyecto y los repositorios de datos identificados, es posible delimitar el alcance del proyecto en un conjunto de objetivos de requisitos que podrán ser luego resueltos a través de la aplicación de procesos de EdI [10].
Sin embargo, y como sucede en otras ingenierías, al comienzo de un proyecto de EdI se tiene la dificultad adicional de no manejar el vocabulario del negocio y de los datos de los miembros de la organización para poder lograr un mejor entendimiento sobre los aspectos del proyecto.
En este sentido, ante la carencia de métodos, técnicas y herramientas asociadas a la ejecución de las tareas relacionadas a la gestión de requerimientos en proyectos de EdI, se ha detectado la necesidad de ofrecer los siguientes elementos:

- un proceso de elicitación de requerimientos para identificar las principales necesidades del cliente, sus expectativas, restricciones y los principales repositorios de datos que son necesarios para realizar el proyecto.
- un conjunto de plantillas que permita documentar todos los requerimientos para que puedan ser, luego, consultados durante la realización del proyecto.
- un proceso de formalización de requerimientos que indique la forma en que se deben completar las plantillas en base a los requerimientos educidos.

3. Solución Propuesta

Para dar respuesta al problema detectado en la sección anterior, se han propuesto varias soluciones (descriptas en la sección 3.1) que buscan dar solución a aspectos puntuales del problema pero que presentan problemas al intentar aplicarlas en un todo. A partir del análisis de estas soluciones parciales existentes se propone su integración en un Modelo de Procesos para la Gestión completa de los Requerimientos en Proyectos de EdI. La propuesta de este modelo se realiza en la sección 3.2.

3.1. Soluciones Existentes

En [11] se realiza una propuesta para la solución del problema definiendo tanto un conjunto de plantillas asociadas al proyecto, los requisitos, los datos y la terminología de la organización así como un proceso general que sirva de guía para la obtención de la información necesaria que debe ser documentada en dichas plantillas. Sin embargo, este proceso se encuentra demasiado vinculado a la documentación de los requerimientos, dejando sin definir las actividades necesarias para educir y entender los requerimientos. De todas formas, ha servido como punto de partida para la definición de las propuestas realizadas a continuación:

1) A partir de la revisión de las plantillas propuestas se ha realizado un ajuste en las mismas con el objetivo de tener un mayor entendimiento y se ha definido un conjunto de normas que indican cómo se deben escribir los requerimientos [12].
2) Se ha definido un proceso de formalización para transformar los requerimientos educidos de forma de poder ser documentados en las plantillas revisadas (a través de la utilización de técnicas de representación de conocimiento provistas por la Ingeniería del Conocimiento [13]).
3) Adaptando el ciclo de vida propuesto por [14] para iniciativas de Data Warehouse & Business Intelligence, se ha propuesto un modelo de procesos para elicitación de requerimientos en proyectos de EdI [15, 16] el cual define un conjunto de fases con las tareas, técnicas y métodos que se deben aplicar, pero que no considera la utilización del proceso de formalización para completar las plantillas correspondientes.

3.2. Modelo de Procesos Propuesto

La solución propuesta consiste en un modelo de procesos que permite relacionar las diferentes propuestas asociadas a la elicitación de requerimientos en proyectos de EdI. El modelo consiste en definir una relación transversal entre las propuestas realizadas en [12], [13], [15] y [16].
El proceso se divide en cuatro fases principales: Definición del Proyecto, Educción del Negocio, Conceptualización del Negocio e Identificación de Procesos de Explotación de Información. Cada fase tiene definido un conjunto de actividades y un conjunto de procesos de formalización asociados a dichas actividades.
La figura 1 muestra el mapa conceptual completo del Modelo de Procesos propuesto, dividido en los diferentes niveles (Fase, Actividad y Proceso de Formalización).

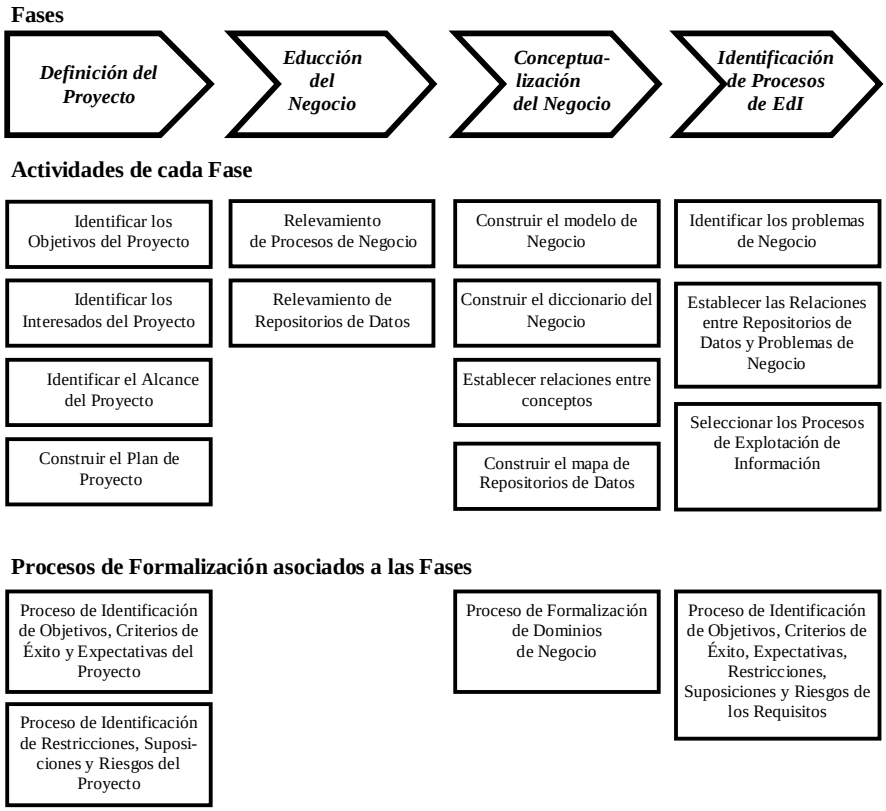


Fig.1 Mapa Conceptual del Modelo de Procesos Propuesto

La fase de Definición del Proyecto tiene por objetivo definir el alcance del proyecto, los interesados y los objetivos que se deben alcanzar. En esta fase se realiza la planificación de actividades de educción de requisitos.

La fase de Educción del Negocio tiene por objetivo comprender el idioma utilizado en el negocio, descubrir las palabras específicas del mismo y cuál es el significado que el negocio le da a esas palabras específicas.
La fase de Conceptualización del Negocio tiene por objetivo definir el negocio en términos de conceptos utilizados en el mismo, vocabulario y repositorios donde se almacena la información de los diferentes procesos del negocio.
La fase de Identificación de Procesos de Explotación de Información define los procesos de minería de datos que se pueden utilizar para resolver los problemas identificados en el proceso de negocio.
En esta instancia, se identifican los roles de las personas que participan en este modelo de procesos. Dichos roles, junto con sus responsabilidades son indicados en la tabla 1.

Rol	Responsabilidades
Líder de Proyecto	Gestionar las acciones para que se lleven a cabo las actividades del proyecto y se cumplan los compromisos del proyecto.
Analista Funcional	Relevar y analizar los diferentes procesos del negocio.
Especialista de Datos	Relevar y analizar las diferentes fuentes de información. Debe tener las capacidades técnicas necesarias para recuperar datos de dichas fuentes.
Analista en Explotación de Información	Establecer relaciones entre datos y relacionar procesos de explotación de información con los problemas del negocio detectados. Aplicar algoritmos de minería de datos sobre los datos relacionados al proceso de negocio para obtener el conocimiento que permita resolver los problemas del negocio detectados.

Tabla 1. Roles del Proceso de Gestión de Requerimientos

3.2.1 Fase de Definición del Proyecto

Durante esta fase se realizan las tareas asociadas a la planificación del proyecto y a establecer el alcance y las personas interesadas en el mismo. Estas actividades son la base de todo proyecto [17]. La figura 2 muestra las actividades y procesos de formalización asociados a esta fase.

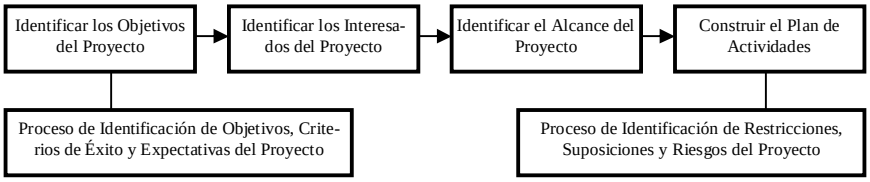


Fig.2 Actividades y Procesos de Formalización de la Fase de Definición del Proyecto

El Líder de Proyecto será el responsable de las diferentes actividades de esta fase. Comienza con la identificación de los objetivos del proyecto, utilizando el Proceso de Identificación de Objetivos, Criterios de Éxito y Expectativas del Proyecto. Con los

objetivos definidos, se identifica la lista de interesados del proyecto (actividad "Identificar los Interesados del Proyecto") en la que figuran quienes permiten definir el alcance del proyecto de EdI, actividad en la que también participa el Analista Funcional. Por último, con el objetivo, alcance e interesados del proyecto definidos, el Lider de Proyecto utiliza el Proceso de Identificación de Restricciones, Suposiciones y Riesgos del Proyecto para construir el Plan de Actividades que servirá como planificación para la ejecución de las diferentes actividades del proceso (actividad de "Construir Plan de Actividades"). Con el plan construido comienzan las actividades asociadas a las siguientes fases.

3.2.2 Fase de Educción del Negocio

Durante esta fase se realizan las tareas que permiten relevar los diferentes procesos de negocio. El Analista Funcional será el responsable de llevar a cabo las actividades definidas para esta Fase. La figura 3 muestra las actividades que la componen.

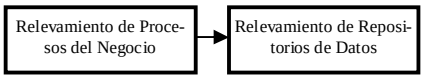


Fig.3 Actividades de la Fase de Educción del Negocio

El Analista Funcional, en la actividad de "Relevamiento de Procesos del Negocio", recopilará la información de los diferentes procesos, mediante la utilización de tecnicas tradicionales de educción, como las que se presentan en [9] y [18]. Esta información documentada será utilizada como referencia en la siguiente fase. Con la ayuda de un Especialista de Datos, deberá identificar los diferentes repositorios de información que existen en la organización (actvidad de "Relevamiento de Repositorios de Datos") y que pueden, o no, estar informatizados.

3.2.3 Fase de Conceptualización del Negocio

En la fase de conceptualización se construye el modelo de negocio que se utiliza como base del Proyecto de EdI. El Analista Funcional será el responsable de las actividades de esta fase que incluyen las actividades definidas en la figura 4.

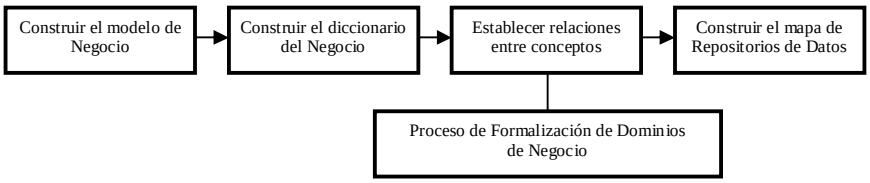


Fig.4 Actividades y Procesos de Formalización de la Fase de Conceptualización del Negocio

Con la información obtenida en la "Educción del Negocio", durante la actividad de "Construir el modelo de Negocio", el Analista Funcional modela los casos de uso de negocio asociados al alcance del proyecto. A partir de este modelo, como parte de la actividad "Construir el Diccionario del Negocio", se analiza el vocabulario utilizado en el negocio y se construye el diccionario de términos asociados a los diferentes procesos de negocio. El Analista de Explotación de Información, en la actividad de "Establecer las Relaciones entre Conceptos", sigue el Proceso de Formalización de Dominios de Negocio, establecido en [13] para generar el Diagrama de Entidad-Relación (DER) asociado a los conceptos. Por último, el rol de Especialista de Datos será el de ser el responsable de relevar las diferentes fuentes de información asociadas a los procesos modelados, las que permitirán establecer las relaciones entre los Casos de Uso identificados y los repositorios de datos existentes. Esta relación se establece en la actividad de "Construir el Mapa de Repositorio de Datos".

3.2.4 Fase de Identificación de Procesos de Explotación de Información

Esta fase es la conclusión del trabajo realizado por el equipo de proyecto y define como resultado final los Procesos de Explotación de Información a utilizar en el proyecto. La figura 5 representa las actividades y los procesos de formalización asociados a esta fase.

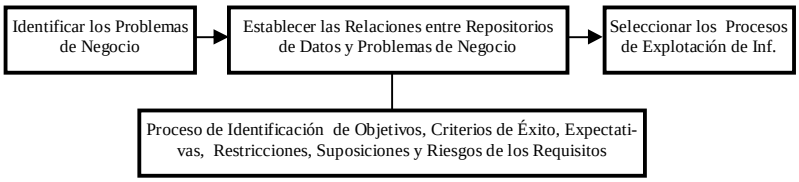


Fig.5 Actividades y Procesos de Formalización de la Fase de Identificación de Procesos de Explotación de Información

El Analista Funcional se encargará de construir una lista que contiene los problemas que el negocio desea resolver (actividad "Identificar los Problemas de Negocio"), y en conjunto con el Analista de Explotación de Información establecerán las relaciones existentes entre los problemas identificados y los repositorios que contienen información útil para la resolución de los problemas (actividad "Establecer las Relaciones entre Repositorios de Datos y Problemas de Negocio) utilizando el Proceso de Proceso de Identificación de Objetivos, Criterios de Éxito, Expectativas, Restricciones, Suposiciones y Riesgos de los Requisitos. Con esta información, y utilizando técnicas propuestas en [15], [16] se podrá/n seleccionar el o los procesos de Explotación de Información que utilizará el proyecto (actividad "Seleccionar Procesos de Explotación de Información").

4. Prueba de Concepto

La prueba de concepto se realizó aplicando el proceso propuesto, al caso de estudio "Detección de Patrones de Daños y Averías" [19].
Como primera fase del proceso se identifican los objetivos que el negocio desea cumplir, los interesados del proyecto y el alcance que se desea alcanzar. Estas tareas se planifican y formalizan en el plan de actividades.
El trabajo a realizar define la lista de objetivos de negocio, entre los que podemos destacar los siguientes:

- OR-01 - Determinar la responsabilidad de siniestralidad en función del tipo de avería y el tipo de transporte.
- OR-02 - Identificar incidentes según el tipo de transporte.
- OR-03 - Identificar tipos de averías y/o daños.
- OR-04- Identificar partes averiadas y/o dañadas que muestren algún tipo de comportamiento.
- OR-05 - Identificar la gravedad de los daños y/o averías.
- OR-06 - Identificar los lugares donde se producen daños y/o averías tratando de definir patrones de comportamiento.
- OR-07 - Determinar en forma estadística: tipos de transporte que producen daños y/o averías como así también, partes, tipos de averías, gravedades, lugares donde se producen.

Durante la fase de "Educción de Negocio" se releva la información asociada a los procesos de negocio y a los repositorios de información asociados. La información relevada puede presentarse en un documento como el presentado en [19] o utilizando cualquier otro mecanismo de documentación tradicional como los indicados en [18].
En la Fase de "Conceptualización del Negocio" se trabaja con los modelos de negocio y de vocabularios, obteniendo el mapa de repositorios y conceptos, utilizando el Proceso de Formalización de Dominios de Negocio para establecer las relaciones entre conceptos. En [13] se presenta el trabajo realizado utilizando este proceso.
Por ultimo, se trabaja en la "Fase de Identificación de Procesos de Explotación de Información". En este caso, podemos identificar los siguientes problemas de negocio:

- Dificultad en identificar daños y averias producidos en unidades automotrices cero kilómetro desde que parten de la fábrica hasta el final del circuito.
- Establecer recursos consumidos por el movimiento de las unidades.
- Elevado costo en la distribución de repuestos para reparar las averías.

En [13] se define cómo establecer las relaciones entre los conceptos y los repositorios de datos. Como última actividad, se deben seleccionar los Procesos de Explotación de Información utilizando la información obtenida durante el proceso, se analizan los diferentes objetivos de requisito identificados y se establecen qué Procesos de Explotación de Información son acordes a cada problema identificado.
De dicho análisis, se puede concluir con las siguientes relaciones entre Procesos de Explotacion de Información y objetivos:

- Para OR1, se selecciona el proceso de descubrimiento de reglas de pertenencia a grupos.

- Para OR2, se selecciona el proceso de descubrimiento de reglas de comportamiento usando “Transporte” como atributo objetivo.
- Para OR3, se selecciona el proceso de descubrimiento de reglas de comportamiento usando “Averia” como atributo objetivo.
- Para OR4, se selecciona el proceso de descubrimiento de reglas de comportamiento usando “Parte” como atributo objetivo.
- Para OR5, se selecciona el proceso de descubrimiento de reglas de comportamiento usando “Gravedad” como atributo objetivo.
- Para OR6, se selecciona el proceso de descubrimiento de reglas de comportamiento usando “Lugar” como atributo objetivo.
- Para OR7, se selecciona el proceso de ponderación de reglas de pertenencia a grupos.

Con este análisis se dan por concluidas las actividades del proceso y se continúa con la ejecución del proyecto de EdI.

5. Conclusiones

Este trabajo presenta una propuesta de Modelo de Procesos para la Gestión de Requerimientos en Proyectos de EdI que permite relacionar diferentes metodologías y herramientas propuestas en otros trabajos y dar un enfoque global a la Gestión de Proyectos de Explotación de Información. El proceso se descompone en cuatro fases, donde se gestiona en forma completa el alcance y los interesados del proyecto (Fase de Definición del Proyecto), se obtiene la información generada en los diferentes procesos de negocio (Educción del Negocio), se modelan los procesos y datos del negocio (Conceptualización del Negocio).
Como futuras líneas de trabajo, se deberán definir o relacionar los procesos de formalización con la fase de Educción de Negocio, y se trabajará en la prueba del proceso en diferentes casos para obtener validaciones empíricas del mismo.

6. Referencias

1. Thomsen, E. (2003). *BI’s Promised Land*. Intelligent Enterprise, 6(4): 21-25.
2. Schiefer, J., Jeng, J., Kapoor, S. & Chowdhary, P. (2004). *Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence*. Proceedings 2004 IEEE International Conference on E-Commerce Technology. Pág. 162-169.
3. Curtis, B., Kellner, M., Over, J. (1992). *Process Modelling*. Communications of the ACM, 35(9): 75-90.
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step Data Mining Guide*. http://tinyurl.com/crispdm. Último acceso Enero de 2013.
5. Pyle, D. (2003). *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers.
6. SAS (2008). *SAS Enterprise Miner: SEMMA*. http://tinyurl.com/semmaSAS. Último acceso Enero de 2013.

7. Vanrell, J., Bertone, R., & García-Martínez, R. (2010). *Modelo de Proceso de Operación para Proyectos de Explotación de Información*. Anales del XVI Congreso Argentino de Ciencias de la Computación, Pág. 674-682. ISBN 978-950-9474-49-9.
8. García Martínez, R. & Britos, P. (2004). *Ingeniería de Sistemas Expertos*. Editorial Nueva Librería.
9. Wiegers, K. (2003). *Software Requirements*. Microsoft Press.
10. Garcia–Martinez, R., Britos, P., Pollo–Cattaneo, F., Rodriguez, D., Pytel, P. (2011). *Information Mining Processes Based on Intelligent Systems*. Proceedings of II International Congress on Computer Science and Informatics (INFONOR–CHILE 2011). Pág. 87–94. ISBN 978–956–7701–03–2.
11. Pollo-Cattaneo, M. F., Britos, P., Pesado, P. & García-Martínez, R. (2010). *Proceso de Educción de Requisitos en Proyectos de Explotación de Información*. En Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica (Editores: R. Aguilar, J. Díaz, G. Gómez, ELeón). Pág. 01-11. Alfaomega Grupo Editor. ISBN 978-607-707-096-2.
12. Deroche, A. & Pollo-Cattaneo, M. F. (2011) *Guía de Buenas Prácticas para Completar las Plantillas de Requerimientos para Proyectos de Explotacion de Información*. Reporte Técnico GEMIS-TD-2011-01-RT-2012-01. Grupo de Estudio de Metodologías para Ingeniería en Software, UTN-FRBA.
13. Vegega, C., Pytel, P., Ramón, H., Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. (2012). *Formalización de Dominios de Negocio para Proyectos de Explotación de Información basada en Técnicas de Ingeniería del Conocimiento*. Proceedings del XVIII Congreso Argentino de Ciencias de la Computación. Pag. 1049-1058. ISBN 978-987-1648-34-4.
14. Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2011). *The data warehouse lifecycle toolkit*. Wiley & Sons.
15. Mansilla, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. (2012). *Modelo de Proceso para Elicitación de Requerimientos en Proyectos de Explotación de Información*. Proceedings Latin American Congress on Requirements Engineering and Software Testing. Pág. 38-45. ISBN 978-958-46-0577-1.
16. Mansilla, D., Pollo, F., Britos, P., García-Martínez, R. (2013). *A Proposal of a Process Model for Requirements Elicitation in Information Mining Projects*. Lecture Notes in Business Information Processing, 139: 165-173. ISBN 978-3-642-36610-9.
17. William, R. 1996. A Guide To The Project Management Body Of Knowledge. PMI Publishing.
18. Sommerville, Ian, Y Peter Sawyer. 1997. Requirements Engineering: A Good Practice Guide. Chichester, England: John Wiley & Sons
19. H. Flores (2009). Detección de Patrones de Daños y Averías en la Industria Automotriz. Tesis de Maestría en Ingeniería en Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.

Efecto de los *trending topics* en el Volumen de Consultas a los Motores de Búsqueda

Santiago Ricci y Gabriel Tolosa
sricci.soft@gmail.com - tolosoft@unlu.edu.ar

Departamento de Ciencias Básicas, Universidad Nacional de Luján

Resumen Las redes sociales se han convertido en aplicaciones muy populares en Internet, principalmente para publicar información y comunicarse en grupos. Una caso bien conocido es Twitter, considerado un servicio de microblogging. Aquí se generan temas que se vuelven muy populares en la red social en un determinado momento, denominados *trending topics* (TT). En este trabajo se intenta determinar el efecto de estos temas en cuanto al volumen de consultas enviadas a un motor de búsqueda web. Para ello, se utiliza información de Twitter y tendencias de búsqueda. Los resultados iniciales muestran indicios de que los *trending topics* se utilizan luego para consultas al buscador: aproximadamente el 65 % de las consultas muestran un aumento del interés cuando son TT e - inclusive - entre el 44 % y 59 % obtienen su pico de interés. Estos resultados se consideran indicios positivos respecto de la hipótesis planteada abriendo oportunidades de aprovechamiento de esta información para optimizar procesos internos de un motor de búsqueda.

Keywords: Twitter, trending topics, motores de búsqueda

1. Introducción

En los últimos cinco años las redes sociales se han convertido en aplicaciones muy populares, principalmente para publicar información y comunicarse en grupos de personas. Mientras que su crecimiento en cantidad de usuarios es exponencial, sus usos son de lo más variado (microblogging, content-sharing, perfiles profesionales, académicas, entre otras). Este tipo de sistema se basa principalmente en la existencia de conexiones "virtuales" entre sus participantes determinadas por el tipo de relación (amigo, seguidor, etc.) y que forman un grafo subyacente. Esta es una diferencia estructural respecto de los hyperlinks de la web donde los enlaces entre objetos son explícitos. La variación en la estructura afecta tanto la forma de obtener reputación en la red y cómo se localiza y disemina la información [10].

Las redes sociales permiten el agregado rápido de contenido y su propagación por el grafo. Uno de los casos bien conocidos es Twitter, una red social que cuenta con millones de usuarios alrededor del mundo [7]. Definida como un servicio de microblogging (ya que sus entradas, llamada *tweets*, tienen un máximo de

140 caracteres), entre sus características más notables se destacan los *trending topics* (TT) o tendencias. Estos resultan de un algoritmo que identifica los temas emergentes más populares[1]. Los TT pueden ser términos, frases o *hashtags*[2] y se encuentran relacionados con los temas más populares en la red social en un determinado momento (por defecto, se determinan de forma personalizada para cada usuario). Es importante destacar, que en [7] se sugiere que gran parte de los TT están relacionados con las noticias del momento.

Por otro lado, se sabe que el buscador de Twitter es usado para monitorear cierto contenido, mientras que los motores de búsqueda Web son empleados para saber más acerca de dicho tema y que muchos usuarios ejecutan la misma consulta tanto en el motor de búsqueda de Twitter como en uno Web con el fin de capturar ambos usos [10]. También, es conocido el uso de Twitter para expresar opiniones acerca de diferentes temas, lo cual se ha traducido en gran cantidad de trabajos que plantean diferentes enfoques sobre cómo realizar minería de opinión sobre la red social [13] [14]. Estas premisas, junto a que según [7], gran parte de los usuarios de Twitter participan en *trending topics*, en conjunto permiten plantear la siguiente hipótesis: **"el hecho de que un tema sea trending topic, se traduce en un aumento en el volumen de consultas relacionadas con dicho tema en los motores de búsqueda Web"**. Si bien parece intuitiva su veracidad, no existen al momento propuestas de metodologías que permitan validarla ni cuantificar tal impacto.

En el presente trabajo se intenta obtener indicios que permitan verificar esta hipótesis. Esta cuestión es importante debido a que los motores de búsqueda deben responder millones de consultas (*queries*) por día, lo cual implica la necesidad de obtener eficiencia y efectividad para poder otorgar a los usuarios respuestas relevantes lo más rápido posible [11]. Entonces, si existe tal relación, puede sacarse provecho de la misma para mejorar el rendimiento a través del uso de técnicas conocidas en el ámbito de los motores de búsqueda por ejemplo, *caching* y *prefetching* de resultados.

Para poder validar definitivamente esta hipótesis se requiere contar con el conjunto de los TT en un periodo dado y la cantidad de consultas que recibió el motor de búsqueda respecto de estos temas en el mismo período. El primero de los conjuntos se puede obtener de forma directa usando la API de Twitter, pero el segundo es propiedad de los proveedores de los servicios de búsqueda y no se encuentra disponible. Para salvar esta situación se propone un método indirecto que – si bien no puede brindar datos absolutos – posibilita obtener indicios concretos acerca de la relación entre los TT y las consultas al motor de búsqueda. Como contribuciones principales, se propone un conjunto de métricas que, combinadas, permiten obtener una caracterización del comportamiento de un TT (mapeado en un *query*) en un período corto de tiempo. Además, se propone un método indirecto para contrastar el comportamiento de un conjunto de consultas antes, durante y después de ser *trending topic*. El uso de métodos indirectos

[1] https://support.twitter.com/articles/101125-about-trending-topics

[2] Cadena de texto precedida por el símbolo # que se utiliza como *keyword* de búsqueda.

es una técnica ampliamente utilizada en el ámbito de la recuperación de información distribuida [8] en casos que los proveedores de información no cooperen con el sistema (por ejemplo, *query-based sampling*). Complementariamente, se plantea un mecanismo para derivar consultas a partir de los TT.

Este artículo extiende trabajos previos [9] con resultados preliminares e incorpora el análisis una mayor cantidad de datos extraídos de la red social. El resto del trabajo se encuentra organizado de la siguiente manera: la sección 2 presenta trabajos relacionados con este estudio. La metodología utilizada, incluyendo las métricas propuestas y los datos analizados se introducen en la sección 3. Los experimentos y resultados se encuentran en la sección 4. Finalmente, se proponen cuestiones para discusión y trabajos futuros.

2. Trabajos relacionados

No es de conocimiento de los autores otros trabajos que traten la hipótesis planteada. Sin embargo, en [7] se realiza un estudio acerca de las características de los TT y se comparan las búsquedas populares ofrecidas por el servicio de Google Trends[3] con los TT, presentando el grado de solapamiento entre ambos, el cual se encontró que es bajo. También se estudió la diferencia en la "frescura" (*freshness*) de los temas en ambos y se halló que en Twitter son más persistentes. Además, se concluye que los usuarios de la red social tienden a hablar sobre noticias y que gran parte (31 %) de los TT duran aproximadamente un día.

En el trabajo de Asur y otros [2] se estudian los TT y se afirma que aquellos con grandes duraciones están caracterizados por la naturaleza "resonante" del contenido de sus *tweets* asociados, el cual proviene, generalmente, de los medios de comunicación tradicionales. De este modo, Twitter se comporta como un amplificador selectivo del contenido generado por los medios tradicionales mediante cadenas de retweets. En [1], se utiliza a los TT como base para predecir los temas que se volverán populares en el futuro cercano. En [4] se estudia al servicio desde el punto de vista estructural y del contenido. Una publicación posterior [6] amplía dicha caracterización incluyendo la distribución geográfica. Aquí se identificaron diferentes clases de usuarios y su comportamiento, junto con patrones de crecimiento y tamaño de la red. Finalmente, en [10] se compara la tarea de búsqueda de los usuarios en Twitter respecto a los motores de búsqueda, pudiendo cuantificar algunas diferencias. Principalmente, hallaron que las consultas a Twitter son mas cortas pero con palabras más largas y una sintaxis más específica. También se usan palabras comunes, se repiten más y cambian menos. Esto se debe a que los usuarios de Twitter habitualmente realizan búsquedas para monitorear nuevo contenido mientras que las búsquedas web se usan para conocer más sobre un tema. Además, los resultados entregados por ambos tipos de servicios son diferentes. Estas observaciones resultan interesantes para el proceso de mapeo entre TT y *queries*.

[3] http://www.google.com/trends/

3. Metodología

El enfoque general se basa en un proceso de tres fases. Como se mencionó, no es posible obtener los datos de TT y queries para un mismo período de tiempo de forma directa. Para salvar la situación se proponen los siguientes pasos:

1. Capturar los *trending topics* de Twitter durante un periodo de tiempo
2. Derivar consultas a partir de los mismos (los TT pueden ser términos, frases o *hashtags* entonces es necesario mapearlos en un *query*)
3. Analizar la evolución de estos queries usando el servicio de Google Trends, aplicando una serie de métricas que intentan capturar su comportamiento.

3.1. Métricas

Para analizar la evolución de las consultas e intentar establecer indicios de que los TT influyen (aumentan) el volumen de consultas a un motor de búsqueda se proponen tres métricas que capturan aspectos diferentes:

1. **Variación del interés (Var):** es la variación porcentual del interés de la consulta derivada para un *trending topic*. Formalmente se define del siguiente modo: sea $I(n, q_{t_i})$ el interés en el *query* q asociado al *trending topic* t_i en el día n (día en que el tema se convierte en *trending topic*), entonces la variación porcentual respecto al día anterior $(n-1)$ está dada por la siguiente ecuación:

$$Var(n, q_{t_i}) = 100\frac{I(n, q_{t_i}) - I(n-1, q_{t_i})}{I(n-1, q_{t_i})} \tag{1}$$

   En el caso que $I(n, q_{t_i}) = 0$ y $I(n-1, q_{t_i}) = 0$, $Var(n-1, q_{t_i}) = 0$; y para el caso que $I(n, q_{t_i}) \neq 0$ y $I(n-1, q_{t_i}) = 0$, $Var(n, q_{t_i}) = 100I(n, q_{t_i})$.
   Esta métrica, intenta capturar el hecho de que si los TT influyen en el volumen de consultas, entonces debe existir una gran diferencia en el interés en dicha consulta entre cuando el tema es TT y cuando no.
2. **Cambio de tendencia (T):** es la cuantificación del cambio en la tendencia que experimenta cierta consulta derivada cuando el tema se convierte en TT, respecto a su tendencia en los siete días anteriores. Formalmente, si q_{t_i} es la consulta derivada para el *trending topic* t_i en el día n, $m_{q_{t_i}}(x, y)$ es la pendiente de la línea de tendencia para la consulta q_{t_i} entre los días x e y, entonces el cambio en la tendencia se define como:

$$T(q_{t_i}) = \frac{m_{q_{t_i}}(n-6, n) - m_{q_{t_i}}(n-7, n-1)}{|m_{q_{t_i}}(n-7, n-1)|} \tag{2}$$

   En consecuencia debe cumplirse que, para una consulta que viene experimentando interés creciente, $T(q_{t_i})$ sea significativa para poder afirmar que el cambio fue consecuencia de que el tema sea *trending topic*.

3. **Detección de picos:** es la aplicación de un algoritmo de detección de picos (*burst detection*) como el presentado en [12]. Dicha publicación plantea que para descubrir regiones con valores "pico" en una serie temporal, debe calcularse la media móvil (MA) y tomar como valores pico aquellos que superen x desvíos estándar sobre el valor medio de MA. Por lo tanto, aquí se lo utilizará con el fin de plasmar que en el día que un tema es *trending topic*, debería registrarse un pico en el interés de su consulta asociada.

3.2. Datos utilizados

Para obtener los TT se utilizó la API REST de Twitter [4] y se consideraron solo los TT para Argentina (no personalizados). Se realizaron dos capturas. La primera, a la que llamaremos C_1 [9], fue realizada entre el 05/12/2012 y el 12/12/2012 (7 días). La segunda, (C_2), fue generada entre el 06/03/2013 y el 20/03/2013 (14 días). En ambas ocasiones, los TT fueron obtenidos a intervalos de 5 minutos aproximadamente (por límites impuestos por el proveedor del servicio). Bajo dichas condiciones, en el caso de C_1 se obtuvieron 2002 muestras de 10 TT. Eliminando los duplicados en todo el período de captura, se obtuvo un total de 573 TT y eliminando solo las repeticiones del mismo día se obtuvo un total de 727 TT. En el caso de C_2, se capturaron 3954 muestras de 10 TT. Eliminando duplicados en todo el período de captura se obtuvieron 956 TT y quitando solo las repeticiones del mismo día se obtuvieron 1422 TT.

3.3. Derivación de consultas

Este aspecto es muy importante en lo que respecta a los objetivos del trabajo. Como se mencionó, la estructura de los TT no es similar a la de las consultas a un motor de búsqueda y para un *treding topic* se pueden derivar varios *queries*. Además, puede no quedar claro qué consulta puede derivarse o las resultantes pueden ser ambiguas (podría emplearse sus *tweets* correspondientes para intentar desambiguar). En el caso de los *hashtags*, puede ser difícil extraer los términos. También es importante tener en cuenta que, según [10], los *queries* realizados en el sistema de búsqueda de Twitter difieren en longitud, en función y en sintaxis respecto a los de un motor de búsqueda (ambos son *free-text queries*, pero en Twitter, '#' y '@' tienen significados especiales). En consecuencia, parece no ser adecuado emplear la consulta asociada a cada TT que devuelve la API de Twitter. Una cuestión importante, también aportada en dicha publicación, es que gran parte (45.95 %) del conjunto intersección entre los queries formulados en Twitter y en la barra de búsqueda de Bing son informacionales y acerca de celebridades.

Teniendo en cuenta lo anterior, se diseñó un procedimiento para derivar las consultas haciendo uso de los siguientes criterios: si el *trending topic* es una frase o un término, entonces la consulta es dicha frase o término. En el caso que sea un *hashtag*, se elimina el # y se intenta separar los términos haciendo uso de

[4] https://dev.twitter.com/docs/api

las mayúsculas (*camelcase*) y los números. Si no se dispone de mayúsculas, se utiliza la primera sugerencia de Google (se consulta al motor de búsqueda) que normalmente obtienen los usuarios a medida que van escribiendo su consulta. Si no existen sugerencias, se ejecuta el TT en Google y se toma la búsqueda sugerida. En caso de no existir tal sugerencia, se toma como consulta el *hashtag* (sin el #).

3.4. Obtención de datos de tendencias de búsqueda

Para analizar los cambios en la popularidad de las consultas derivadas se requiere alguna fuente de información que disponga de los datos correspondientes a los queries derivados de los TT. Dado que no es posible tener acceso al log de consultas de un buscador web (menos aún en un período puntual), se optó por utilizar los datos del servicio Google Trends. Este servicio permite ejecutar una consulta y obtener la evolución del interés a lo largo del tiempo fijando una ventana temporal. Si bien la documentación[5] de la herramienta no especifica el método exacto para calcular los valores que devuelve, se sabe que los datos reflejan el número de búsquedas de un término en comparación al total de búsquedas realizadas en Google a lo largo del tiempo[6]. Además, los datos están normalizados y son presentados en una escala de 0 a 100, donde el valor 0 se corresponde a la falta de datos (el servicio solo muestra los datos de los términos que sobrepasen un cierto límite en los volúmenes de búsqueda).

Luego, para cada *query* derivado de cada TT del conjunto de cada día, tanto de C_1 como de C_2, se consultó la información empleando una ventana temporal de 30 días y el filtro de datos para Argentina[7]. Según la magnitud del volumen de búsquedas, la cantidad de puntos retornados por la herramienta puede variar. Para C_1, 314 TT (43 % del total) obtuvieron 29 observaciones (puntos). En el caso de C_2, el 55 % (788 TT) obtuvo 26 puntos. En ambos casos, la cantidad de observaciones retornadas por las consultas restantes no se consideran suficientes para los experimentos que aquí se proponen.

4. Experimentos & Resultados

4.1. Variación del Interés

Para cada TT se calculó la primera métrica (Ecuación 1). Luego, se calculó la cantidad de consultas que tuvieron interés creciente, decreciente y constante. La Tabla 1 resume los resultados obtenidos tanto para C_1 como para C_2 . De allí, es importante destacar, que en ambos casos más del 65 % muestra un interés creciente. En la Tabla 2 se muestra el detalle. Se puede apreciar que en aproximadamente la mitad de los casos, el crecimiento fue del 50 % o superior.

[5] http://support.google.com/trends/?hl=es#topic=13762

[6] http://support.google.com/trends/answer/87285?hl=es&ref_topic=13975

[7] Para poder cumplir con los términos del uso del servicio se solicitó colaboración a un grupo de estudiantes entre quienes se repartió la tarea.

Tipo de variación	Porcentaje	
	C_1	C_2
Positivo	71,97 %	67,01 %
Negativo	19,43 %	27,92 %
Nulo	8,60 %	5,08 %

Tabla 1. Variación en el interés respecto al día anterior.

En la figura 1 se muestra el crecimiento acumulado para ambas capturas, las cuales presentan comportamientos similares: alrededor del 30 % del total de *trending topics* con interés respecto al día anterior creciente, obtuvo un aumento de más del 100 %. Sin embargo, pese a que un buen porcentaje de los TT obtiene un aumento significativo, hay que observar la tendencia que venían experimentando las consultas en una ventana temporal mayor. Si la misma es creciente y pronunciada, el hecho de que el tema sea tendencia no posee un gran efecto sobre la popularidad de la consulta.

. Porcentaje de crecimiento (x)	Porcentaje con $Var(n, q_{t_i}) > x$	
	C_1	C_2
10	13,72 %	18,94 %
20	10,18 %	10,04 %
30	13,27 %	10,04 %
40	6,64 %	5,87 %
50	4,42 %	7,58 %
60	5,75 %	6,44 %
70	4,87 %	3,41 %
80	5,75 %	3,60 %
90	0,44 %	3,03 %
100	5,75 %	1,70 %
200	10,18 %	10,23 %
300	4,42 %	4,17 %
400	2,21 %	2,27 %
500	3,54 %	0,95 %
>500	8,86 %	11,73 %

Tabla 2. Detalle del aumento del interés respecto al día anterior.

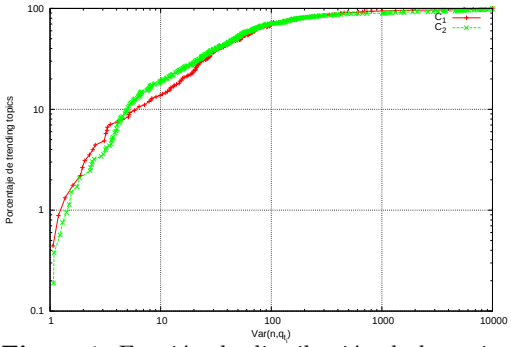


Figura 1. Función de distribución de la variación en el interés respecto al día anterior.

4.2. Cambio de tendencia de las consultas

Para cada consulta derivada se calculó la línea de tendencia de los 7 días anteriores a que el tema fuera TT (tanto en C_1 como en C_2). En la Tabla 3 se presentan los resultados. Para ambas capturas se observa que más del 50 % de los TT analizados posee una tendencia creciente. Esto plantea dos cuestiones: por un lado, pueden existir casos en los que la consulta derivada venía experimentando una tendencia decreciente, pero el interés aumentó al convertirse su tema asociado en TT. Por el otro, al menos que el cambio en la popularidad sea lo "suficientemente grande" no puede considerarse como válida la hipótesis. Por consiguiente, se elaboró la Figura 2 que relaciona a $Var(n, q_{t_i})$ con la pendiente de la línea de tendencia. El área recuadrada es la que debe observarse, dado que idealmente, la nube de puntos debería estar ubicada cerca del eje de las abscisas y alejada del eje de las ordenadas. Esto implica, que son de interés aquellas

consultas derivadas de los TT que venían experimentando popularidad relativamente constante hasta que el tema se convirtió en TT, momento en el cual se produce un aumento significativo en el interés. Obsérvese que en la figura, la nube de puntos está centrada en torno a $Var(n, q_{t_i}) \approx 80$ para ambas capturas y que la mayor densidad de puntos se encuentra en las pendientes con valores pertenecientes al intervalo $[-5; 5]$. Por lo tanto, existen casos donde se observa una tendencia decreciente en el interés y el día en que el tema se convirtió en TT, dicho interés creció abruptamente.

Tendencia	Porcentaje	
	C_1	C_2
Creciente	61,46 %	56,09 %
Decreciente	34,71 %	37,56 %
Estacionaria	3,82 %	6,35 %

Tabla 3. Tendencias de los 7 días previos a que un tema sea *trending topic*.

Observando específicamente las pendientes positivas, se encuentran casos en los cuales la variación porcentual de interés dada por la primer métrica es grande en comparación al crecimiento que debería esperarse dada la pendiente de la línea de tendencia y viceversa. Es importante aclarar que la tendencia que experimenta una consulta en los últimos 7 días no se corresponde necesariamente con la tendencia considerando una ventana temporal mayor. Por ejemplo, podría darse el caso de un *query* cuya popularidad evolucione de forma cíclica, con un período mayor a la ventana temporal actualmente considerada. El análisis de estos casos se deja para trabajos futuros.

Los resultados de la segunda métrica se muestran en la Tabla 4. Para C_1, en aproximadamente en el 40 % de los casos, el crecimiento de la pendiente es de al menos dos veces la pendiente de la recta que no considera el día en el que el tema fue *trending topic*. En cambio para C_2, en el 33 % de los casos aproximadamente $T(q_{t_i}) > 2$. Nuevamente, la ventana temporal empleada puede influir en los resultados. Si bien la figura no permite afirmar que la hipótesis se cumple para todos los casos, muestra que existen casos para los que pareciera que sí se cumple. Por otro lado, el análisis del cambio en la pendiente plasmado por la segunda métrica parece reforzar lo previamente dicho.

$T(q_{t_i})$	Porcentaje	
	C_1	C_2
$<$ -0.5	5,31 %	12,12 %
< 0	3,98 %	6,63 %
< 0.5	11,50 %	14,20 %
< 1	15,49 %	15,34 %
< 1.5	12,83 %	8,52 %
< 2	10,62 %	10,04 %
< 2.5	4,87 %	3,60 %
< 3	3,98 %	3,41 %
< 5	9,29 %	7,95 %
< 10	9,73 %	9,47 %
≥ 10	12,39 %	8,71 %

Tabla 4. Relación entre pendientes.

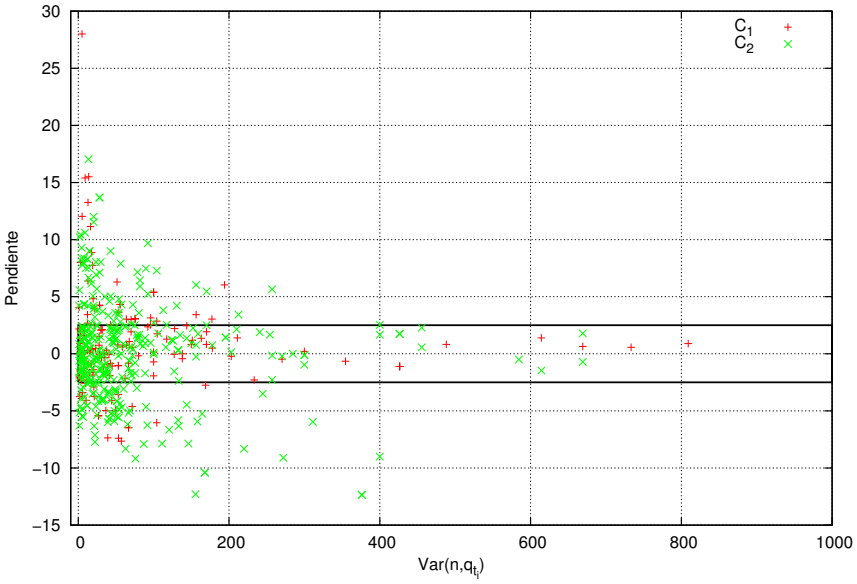


Figura 2. Relación entre la pendiente de la línea de tendencia del interés en la consulta en los 7 días anteriores a que el tema sea TT y la métrica determinada por la Ecuación 1.

4.3. Detección de picos

Por último se ejecutó el algoritmo de detección de picos sobre C_1 y C_2, utilizando la configuración recomendada en [12], es decir, una ventana de 7 días y $x = 1{,}5$. En el caso de C_1, del total de TT para los que se cuenta con suficientes datos, 162 (51,6 %) tuvieron un pico en el interés el día que fueron TT. En el caso de C_2, 297 (37,69 %). La Figura 3 muestra el caso del *trending topic* "Papa Francisco", del cual se derivó el *query* "papa francisco". Esta consulta tuvo un pico en el interés el día en que su tema asociado fue *trending topic*. La Figura 4 muestra un caso donde no se identificó un pico de interés en la consulta. En ambas gráficas puede observarse la evolución de la media móvil hasta el día en que el tema fue *trending topic* y el umbral (*cutoff*) para considerar un pico en el interés.

Es curioso observar que, en el caso de la Figura 4, el pico se produce un día después que el tema fue *trending topic*. En realidad, el tema apareció como TT aproximadamente a las 22 hs. (22:09:38 es la hora exacta en la que aparece el primer registro), con lo cual se puede suponer que parte del tráfico va a estar contenido en el día siguiente. Por lo tanto, se procedió a analizar también el día siguiente. Los resultados muestran que la cantidad de consultas que obtienen un pico de interés se incrementa en un 7,6 % para C_1 y un 6,7 % para C_2. Esto sugiere que debe estudiarse la duración de los TT. [7] y [9] presentan resultados de dicho estudio. Teniendo esto presente, se hace evidente la necesidad de estudiar

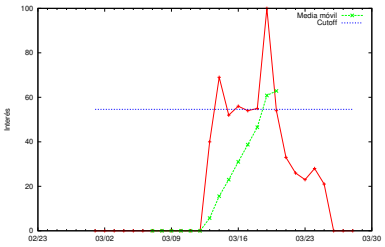


Figura 3. *Trending topic* de C_2 cuyo *query* obtuvo un pico en el interés.

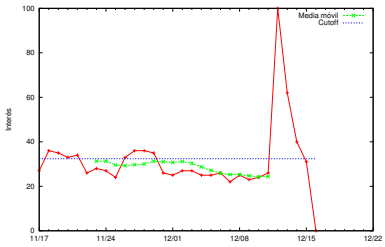


Figura 4. *Trending topic* de C_1 cuyo *query* no obtuvo un pico en el interés.

el aumento del tráfico en relación a la duración del TT, cuestión que se deja abierta para trabajos futuros. Esto es importante, dado que debe verificarse que la duración de un *trending topic* sea suficiente como para generar el tráfico necesario para alterar de forma abrupta la tendencia que venía experimentando la consulta.

5. Discusión y Trabajos Futuros

En este trabajo se intenta determinar el efecto de la aparición de información de tendencias en las redes sociales respecto de las consultas enviadas a un motor de búsqueda web. Para ello, se utilizan los *trending topics* de la red Twitter y la información de tendencias de búsqueda provista por Google como método indirecto, ya que no existe la posibilidad de obtener los archivos log de las consultas. Se proponen tres métricas que intentan capturar diferentes aspectos: variación porcentual, tendencia y picos. Los resultados presentan indicios de que los *trending topics* se utilizan luego para consultas al buscador. En particular, aproximadamente el 65 % de las consultas muestran un aumento del interés (tendencia) cuando son TT e - inclusive - entre el 44 % y 59 % obtienen el pico de interés en ese mismo día o el siguiente. Estos resultados se consideran indicios positivos respecto de la hipótesis planteada aunque al momento y con los datos disponibles no se puede cuantificar.

Entonces, se abre la posibilidad de varios trabajos futuros: por un lado, seguir desarrollando métodos indirectos para obtener más información de otros sistemas similares y poder realizar mediciones con mayor grado de precisión. Esto incluye diferentes enfoques (por ej. semánticos) para obtener las consultas relacionadas a un TT. Además, debe estudiarse qué sucede luego de finalizado el período en el cual un tema es *trending topic*. Por otro lado, explotar esta relación para optimizar dos de los mecanismos principales utilizados por los servicios de búsqueda de gran escala como el caching y el prefetching de resultados. De esta manera, se podrían *anticipar* peticiones en momentos de menor actividad y disminuir las posibilidades de sobrecarga.

Referencias

1. Agarwal, S.; Agarwal, S. Social networks as Internet barometers for optimizing content delivery networks. In 3rd International Symposium on Advanced Networks and Telecommunication Systems (ANTS). 2009.
2. Asur, S.; Huberman, B. A.; Szabo, G.; Wang, C. Trends in social media: Persistence and decay. In 5th International AAAI Conference on Weblogs and Social Media. 2011.
3. Callan, J.; Connel, M. Query-based sampling of text databases. In ACM Transactions on Information Systems, v. 19, n. 2, pp. 97-130. 2001.
4. Java, A.; Song, X.; Finin, T.; Tseng B. Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD '07), pp. 56-65. 200
5. Jonassen, S.; Barla Cambazoglu B.; Silvestri F. Prefetching query results and its impact on search engines. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12), pp. 631-640. 2012.
6. Krishnamurthy, B.; Gill, P.; Arlitt M. A few chirps about twitter. In Proceedings of the first Workshop on Online Social Networks (WOSN '08), pp. 19-24. 2008.
7. Kwak, H.; Lee, C.; Park, H.; Moon, S. What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World Wide Web (WWW '10), pp. 591-600. 2010.
8. Luo S. Federated search of text search engines in uncooperative environments. PhD Thesis. Carnegie Mellon University. 2006.
9. Ricci, S.Impacto de las Redes Sociales en la Popularidad de las Consultas a Motores de Búsquedas. Jornadas Argentinas de Informatica. JAIIO 2013.
10. Teevan, J.; Ramage, D.; Ringel Morris, M. #TwitterSearch: a comparison of microblog search and web search. In Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM '11), pp. 35-44. 2011.
11. Tolosa G. H.; Feuerstein E. Mejoras algorítmicas y estructuras de datos para búsquedas altamente eficientes. En Proceedings del XIV Workshop de Investigadores en Ciencias de la Computación (WICC 2012), p. 740-744. 978-950-766-082-5. 2012.
12. Vlachos, M.; Meek, C.; Vagena, Z.; Gunopulos, D. Identifying similarities, periodicities and bursts for online search queries. In Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD '04), pp. 131-142. 2004.
13. Wang, X.; Wei, F.; Liu, X.; Zhou, M.; Zhang, M. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In Proceedings of the 20th ACM international conference on Information and Knowledge Management (CIKM '11) pp. 1031-1040. 2011.
14. Xinfan M.; Furu W.; Xiaohua L.; Ming Z.; Sujian L.; Houfeng W. Entity-centric topic-oriented opinion summarization in twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD '12), pp. 379-387. 2012.

Propuesta de Métricas para Proyectos de Explotación de Información

Diego Basso[1,2], Darío Rodríguez[3], Ramón García-Martínez[3]

1. Grupo de Investigación en Ingeniería. Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza
2. Programa de Maestría en Ingeniería de Sistemas de Información. UTN-FRBA
3. Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información. Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús.
diebasso@yahoo.com.ar, drodrigu@unla.edu.ar, rgm1960@yahoo.com

Resumen. Los proyectos de explotación de información requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Las métricas usuales para realizar una estimación no se consideran adecuadas ya que los parámetros a ser utilizados son de naturaleza diferentes y no se ajustan a sus características particulares. En este contexto, se plantea una propuesta de métricas aplicables al proceso de desarrollo de Proyectos de Explotación de Información, siguiendo los lineamientos del Modelo de Procesos para Proyectos de Explotación de Información para PyMEs.

Palabras claves: Métricas. Explotación de Información. Modelo de Procesos

1. Introducción

En la Ingeniería de Software, los proyectos de desarrollo tradicionales aplican una amplia diversidad de métricas e indicadores a distintos atributos y características de los productos y procesos del desarrollo, con el fin de garantizar la calidad del software construido. En el ámbito de la Ingeniería en Conocimiento, especialmente en el desarrollo de sistemas expertos o sistemas basados en conocimientos, mediante la medición de la conceptualización se puede estimar actividades futuras y obtener información del estado de madurez del conocimiento sobre el dominio y sus particularidades [Hauge et al., 2006]. Estas métricas de madurez de conceptualización para Sistemas Expertos aplicadas en [Pollo-Cattaneo, 2007] brindan además información sobre la complejidad del dominio. Los Proyectos de Explotación de Información también requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Sin embargo, dada las diferencias que existen con un proyecto clásico de construcción de software, las métricas usuales para realizar una estimación no se consideran completamente adecuadas ya que los parámetros a ser utilizados son de naturaleza diferentes y no se ajustan a sus características particulares.

Este trabajo tiene como objetivo presentar una propuesta de métricas aplicables al proceso de desarrollo de proyectos de Explotación de Información propuesto en [Vanrell, 2011] para PyMEs. Para ello, primero se realiza una introducción a las características del proceso de desarrollo mencionado, los procesos de explotación de información basados en sistemas inteligentes y la categorización definida en el método DMCoMo para agrupar dichas características (sección 2); luego se delimita el problema (sección 3) presentando una propuesta de solución (sección 4) y las consideraciones para la validación de la propuesta (sección 5), finalizando con la puntualización de algunas conclusiones parciales (sección 6).

2. Estado de la Cuestión

En esta sección se introduce el proceso de desarrollo para proyectos de Explotación de Información (sección 2.1), los procesos de explotación de información basados en sistemas inteligentes (sección 2.2) y el método de estimación desarrollado para este tipo de proyectos - DMCoMo (sección 2.3).

2.1. Proceso de desarrollo para proyectos de Explotación de Información

Las etapas de desarrollo de los proyectos de Explotación de Información no coinciden naturalmente con las etapas mediante las cuales se desarrollan los proyectos de software tradicionales. El modelo de procesos para proyectos de Explotación de Información propuesto en [Vanrell, 2011] plantea dos procesos principales: uno vinculado a la administración de proyectos de explotación de información y otro relacionado con el desarrollo del proyecto. Para el interés de este trabajo, nos centraremos en el segundo de los procesos mencionados, cuyos subprocesos y tareas están definidas a partir de las fases de desarrollo planteadas por la metodología CRISP-DM. Estos subprocesos son: Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Entrega. Claramente estos subprocesos difieren de las etapas definidas para un proyecto de desarrollo de software tradicional (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre).

2.2. Procesos de Explotación de Información basados en sistemas inteligentes

En el trabajo realizado por [Britos, 2008] se proponen cinco procesos de explotación de información que podrían aplicarse a la etapa de Modelado del proceso de desarrollo propuesto por [Vanrell, 2011]. Estos procesos son:

- Proceso de Descubrimiento de Reglas: permite identificar condiciones para obtener resultados del dominio del problema.
- Proceso de Descubrimiento de Grupos: permite identificar una partición dentro de la información disponible dentro del dominio de un problema.

- Proceso de Ponderación de Interdependencia de Atributos: se utiliza cuando se desea identificar los factores con mayor incidencia sobre un determinado resultado de un problema.
- Proceso de Descubrimiento de Reglas de Pertenencia a Grupos: permite identificar las condiciones de pertenencia a cada una de las clases en una partición desconocida pero que se encuentra presente en la masa de información disponible sobre el dominio del problema.
- Proceso de Ponderación de Reglas de Comportamiento o de Pertenencia a Grupos: se utiliza cuando se requiere identificar las condiciones con mayor incidencia sobre la obtención de un determinado resultado en el dominio del problema, ya sea por la mayor medida en la que inciden sobre su comportamiento o las que mejor definen la pertenencia a un grupo.

A su vez, entre las tecnologías de sistemas inteligentes aplicadas a la explotación de información [García Martínez et al., 2003] se encuentran: los algoritmos de inducción o TDIDT, los mapas auto organizados de Kohonen o SOM (Self Organized Maps) y las redes bayesianas [Britos, 2008].

2.3. Método de Estimación para Proyectos de Explotación de Información (DMCoMo)

En [Marbán, 2003] se define un método analítico de estimación para proyectos de explotación de información el cual se denomina "Matemático Paramétrico de Estimación para Proyectos de Data Mining" (en inglés Data Mining COst MOdel, o DMCoMo). Este método, validado y desarrollado a través de una herramienta de software en [Pytel, 2011], permite estimar los meses/hombre que serán necesarios para desarrollar un proyecto de explotación de información desde su concepción hasta su puesta en marcha. Para realizar la estimación se definen seis categorías [Marbán, 2003] para vincular las características más importantes de los proyectos de explotación de información. Estas categorías son las siguientes: Datos, Modelos, Plataforma, Técnicas y Herramientas, Proyecto y Personal (Staff del Proyecto).

3. Definición del Problema

Al igual que en los proyectos de desarrollo de software tradicionales, los proyectos de Explotación de Información requieren de un proceso de planificación que permita estimar sus tiempos y medir el avance del producto en cada etapa de su desarrollo y calidad del mismo. Sin embargo, dada las diferencias que existen entre un proyecto clásico de construcción de software y un proyecto de explotación de información, las métricas usuales para realizar una estimación no se consideran completamente adecuadas ya que los parámetros a ser utilizados son de naturaleza diferentes y no se ajustan a sus características particulares, por ejemplo cantidad de fuentes de información, nivel de integración de los datos, el tipo de problema a ser resueltos, entre las más representativas de este tipo de proyectos. Como se mencionó en la sección anterior, el modelo de procesos para proyectos de Explotación de Información

propuesto en [Vanrell, 2011] plantea dos procesos principales: el proceso de administración de proyectos y el de desarrollo de proyectos de explotación de información. El proceso de administración de proyectos, se encarga de recolectar información necesaria para aumentar la calidad del proceso de desarrollo permitiendo realizar ajustes en el mismo y mantener un estándar en la realización de proyectos. Sin embargo, no plantea qué métricas utilizar para evaluar la calidad del proceso de desarrollo en este tipo de proyectos.

4. Solución Propuesta

En base a la categorización que en [Marbán, 2003] se realiza sobre el modelo de estimación DMCoMo, se plantea la utilización de aquellas categorías que sean aplicables al proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011] como forma de clasificación de las métricas propuestas, focalizado este proceso en proyectos pequeños [Pytel, 2011] que son los que usualmente requieren las PyMEs. Estas métricas a su vez, se orientan a procesos de explotación de información que utilizan tecnologías de sistemas inteligentes.

4.1. Métricas de Datos

A partir del proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011], se han considerado métricas de datos para las tareas que se indican en la tabla 1.

Tabla 1. Propuesta de Métricas de Datos

Subproceso: Entendimiento de los Datos
Tarea: Reunir los datos iniciales
Métricas Propuestas
▪ NT = Número total de fuentes de datos (tablas) necesarias para el proyecto. Se incluyen tablas internas y externas. ▪ NA (T)[1] = Número de atributos en la tabla T. ▪ NTA = Número total de atributos de las tablas. $NTA = \sum_{1}^{i} NA(T_i)$ ▪ NR (T)[1, 2]= Número de registros de la tabla T. Nota [1] – Métrica utilizada en la tarea: Explorar los datos y Limpiar los datos
Tarea: Explorar los datos
Métricas Propuestas
▪ NVN (T) = Número de valores nulos o faltantes en la tabla T. ▪ NANR (T) = Número de atributos nulos o faltantes del registro R en la tabla T. ▪ NCT (T) = Nivel de compleción de la tabla T. Mide el grado de compleción que tiene la tabla. $NCT(T) = 1 - \frac{NVN(T)}{NR(T)}$ ▪ DANR (T) = Densidad de atributos nulos o faltantes de un registro en la tabla T. Mide la proporción de atributos nulos o faltantes que tiene un registro R en la tabla T. $DANR(T) = \frac{NANR(T)}{NA(T)}$
Tarea: Verificar la calidad de los datos
Métricas Propuestas
▪ NAN (T) = Número de atributos a normalizar (transformados para su utilización) en la tabla T. ▪ NVE (A) = Número de valores erróneos del atributo A en la tabla. ▪ GIA (A) = Grado de integridad de un atributo A. $GIA(A) = 1 - \frac{NVE(A)}{NR(A)}$ donde NR es el número de registros que tiene el atributo A.

- NRN (T) = Número de registros con atributos nulos o faltantes en la tabla T.
- NRDE (T) = Número de registros con valores de atributos erróneos en la tabla T. Mide el nivel de precisión en los datos.
- NRVD (T) = Número de registros con valores duplicados en la tabla T.
- NRD (T) = Número de registros defectuosos en la tabla T. $NRD(T) = NRN(T) + NRDE(T) + NRVD(T)$

- GVRT (T) = Grado de validación de los registros de la tabla T. Mide el porcentaje de registros válidos obtenidos. Se recomienda que la validación obtenida sea al menos de 75%. $GVRT(T) = \frac{NR(T) - NRD(T)}{NR(T)} * 100$
- NVD = Nivel de volatilidad de los datos. Mide la frecuencia con que se cambian los datos en el tiempo [*0: no varían – 1: baja volatilidad – 2: volatilidad media – 3: alta volatilidad - 4: gran volatilidad*]

Subproceso: Preparación de los Datos

Tarea: Seleccionar los datos

Métricas Propuestas

- NAU (T) = Número de atributos útiles (significativos para el proyecto) y que no necesitan modificarse en la tabla T.
- NAM (T)[3] = Número de atributos útiles que se deben modificar (se incluye los atributos a normalizar) en la tabla T.

- Nota [3] – Métrica utilizada en la tarea: Limpiar los datos

Tarea Limpiar los datos

Métricas Propuestas

- NRE (T) = Número de registros eliminados en la tabla T.
- NAE (T) = Número de atributos a eliminar (no significativos para el proyecto) en la tabla T.
- GUT (T)[4] = Grado de utilidad de la tabla T. Mide el porcentaje de atributos útiles de la tabla T para el proyecto. A cada atributo se le asigna un peso según su estado de utilidad (*útiles = 0 / modificados = 0,5 / eliminados = 1*)

$$GUT(T) = \frac{NA(T) - (NAE(T) + 0{,}5 * NAM(T))}{NA(T)} * 100$$

Nota [4] – Métrica utilizada en la tarea: Integrar los datos

Tarea: Construir los datos

Métricas Propuestas

- NANI (TI) = Número de atributos nuevos a agregar en la integración de una única tabla TI.

Tarea: Integrar los datos

Métricas Propuestas

- NR (TI) = Número de registros de la tabla integrada.
- GUTAP = Grado de utilidad total de los atributos para el proyecto. Esta métrica mide el nivel de integración de todos los atributos disponibles en una única tabla. $GUTAP = \sum_{1}^{i} GUT(T_i)$

Si 0 <GUTAP < 40% los atributos no son usables.
Si 41% <GUTAP< 80% los atributos son aceptablemente usables.
Si 81 <GUTAP< 100% los atributos son muy usables.

4.2. Métricas de Modelos

En los proyectos de Explotación de Información es necesario evaluar la calidad de los modelos obtenidos de la manera más precisa que sea posible, para garantizar la aplicación de los mismos. Al no existir un modelo mejor que otro de manera general, para cada problema nuevo es necesario determinar con cuál se pueden obtener mejores resultados. A partir de los procesos de explotación de información definidos en [Britos, 2008] y según su tarea de descubrimiento, se pueden clasificar los modelos en: Descubrimiento de Grupos, Descubrimiento de Reglas y Descubrimiento de Dependencias Significativas.

A continuación se realiza una breve descripción de cada uno de estos modelos, las técnicas de sistemas inteligentes aplicables y se presentan los criterios escogidos para la evaluación de los modelos.

Descubrimiento de Grupos: Tiene por objetivo la separación de los datos en grupos (clusters) o clases basándose en la similitud de los valores de sus atributos. Todos los elementos del grupo deben tener características comunes pero a su vez entre los grupos los objetos deben ser diferentes [Britos, 2008]. Dentro de las

tecnologías inteligentes apropiadas para realizar agrupamiento están los mapas auto organizados de Kohonen (SOM – Self Organized Map, por sus siglas en inglés). Al construir un modelo de agrupamiento basado en mapas auto organizados, se define el número de grupos a priori, generando la necesidad de evaluar diferentes topologías para escoger de entre todas la mejor sub-óptima para la solución del problema. Estos mapas se basan en el aprendizaje no supervisado y competitivo. El factor de calidad del modelo generado está basado en el número de grupos que se definen al inicio, ya que al establecerse anticipadamente puede limitar la calidad de agrupamiento del algoritmo, y al ser una tarea de análisis exploratorio, no se sabe con precisión cuantos grupos pueden contener los datos.

Descubrimiento de Reglas: Es uno de los modelos más importantes de de la explotación de información. Se utiliza para encontrar las reglas de clasificación de un conjunto de elementos con base en los valores de sus atributos. El objetivo es lograr modelos de clasificación (expresados en reglas) que determinen correctamente la clase ante elementos no previstos anteriormente [Britos, 2008]. Los algoritmos mayormente utilizados para las tareas de clasificación son los algoritmos de inducción TDIDT (ID3, C4.5 y C5). Para evaluar un modelo de clasificación existen diversas métricas, sin embargo no es aconsejable emplear una sola de ellas ya que es común que una técnica de clasificación presente buenos resultados en una métrica y malos en otra. Por otra parte, al momento de aplicar las técnicas de clasificación se debe tener en cuenta cómo están distribuidos los elementos respecto a la clase o cluster. Puede ocurrir que al no estar balanceadas las clases los clasificadores estén sesgados a predecir un porcentaje más elevado de la clase más favorecida. Las métricas propuestas para evaluar un modelo de clasificación estarán basadas en la matriz de confusión que se obtiene cuando se prueba el clasificador en un conjunto de datos que no intervienen en el entrenamiento. Una matriz de confusión permite conocer la distribución del error a lo largo de las clases o clusters, cuando se prueba un clasificador en un conjunto de datos que no intervienen en el entrenamiento. Una matriz de confusión general tiene la siguiente estructura:

		Clase (cluster) predicha		
		Si	No	Totales
Clase (cluster) real	Si	Verdaderos Positivos (VP)	Falsos Negativos (FN)	Total Positivos Reales (TPR)
	No	Falsos Positivos (FP)	Verdaderos Negativos (VN)	Total Negativos Reales (TNR)

Los valores que se encuentran a lo largo de la diagonal principal de la matriz, representan las clasificaciones correctas y los que están a lo largo de la diagonal secundaria representan los errores (la confusión) entre las clases.

Descubrimiento de Dependencias Significativas: Consiste en encontrar modelos que describan dependencias o asociaciones significativas entre los datos. Las dependencias pueden ser usadas como valores de predicción de un dato, teniendo información de los otros datos. El análisis de dependencias tiene relación con la clasificación y la predicción, donde las dependencias están implícitamente usadas para la formulación de modelos predictivos [Britos, 2008]. Dentro de las técnicas de sistemas inteligentes apropiadas para realizar análisis de dependencias se encuentran las Redes Bayesianas. Si bien las redes bayesianas pueden utilizarse

dentro de los modelos de clasificación, hasta el momento no se han encontrado métricas significativas que establezcan un criterio adecuado de evaluación de dependencias ni de ponderación de atributos significativos.

Los modelos descriptos se corresponden con procesos de explotación de información basados en tecnologías de sistemas inteligentes unitarias [Britos, 2008]. En el caso del proceso de Descubrimiento de Reglas de Pertenencia a Grupos se necesita aplicar una combinación de los modelos de Descubrimiento de Grupos y Descubrimiento de Reglas; y en el caso del proceso de Ponderación de Reglas de Comportamiento o de Pertenencia a Grupos una combinación de los modelos de Descubrimiento de Grupos, Descubrimiento de Reglas y Descubrimiento de Dependencias Significativas, respectivamente. A partir del proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011] y de la clasificación de los modelos establecida anteriormente, se han considerado métricas de modelos para las tareas que se indican en la tabla 2.

Tabla 2. Propuesta de Métricas de Modelos

Subproceso: Modelado
Tarea: Seleccionar técnica de modelado
Métricas Propuestas
▪ NM = Número de modelos a construir para el proyecto. ▪ NE (M) = Número de elementos (registros o casos) en el modelo M. ▪ NA (M) = Número de atributos en el modelo M. - PAN (M) = Porcentaje de atributos numéricos en el modelo M. - PANN (M) = Porcentaje de atributos no numéricos en el modelo M.
Tarea: Generar el diseño del test
Métricas Propuestas
▪ NEE (M) [5]= Número de elementos a utilizar para el entrenamiento del modelo M. ▪ NEP (M) [5] = Número de elementos a utilizar para las pruebas del modelo M. Nota [5] – Métrica utilizada en la tarea: Evaluar el modelo
Tarea: Construir el modelo
Métricas Propuestas
▪ NMDM (M) [6]= Número de modelos de explotación de información aplicados para construir el modelo M. Los modelos pueden ser: descubrimiento de grupos, descubrimiento de reglas y descubrimiento de dependencias significativas. Esta métrica tomará el valor 1, 2 ó 3, según la cantidad de modelos aplicados. Nota [6] – Métrica utilizada en la tarea: Evaluar los resultados
Tarea: Evaluar el modelo
Métricas Propuestas
o Modelo de Descubrimiento de Grupos ▪ NC (M) = Número de agrupamientos (cluster o clases) generados en el modelo M. ▪ NR (C) = Número de elementos agrupados en el cluster C. ▪ PR (C) = Porcentaje de elementos agrupados en el cluster C, respecto del total. ▪ NRP (C) = Número de elementos utilizados para las pruebas del cluster C. ▪ NEPE (M) = Número de elementos del conjunto de prueba del modelo M que fueron incorrectamente agrupados. ▪ TEMA (M) = Tasa de error del modelo de agrupamiento o clustering. Mide el porcentaje de elementos de prueba que fueron mal agrupados. $TEMA(M) = \frac{NEPE(M)}{NEE(M) + NEP(M)} * 100$ ▪ EMA (M) = Nivel de exactitud del modelo de agrupamiento. $EMA(M) = 1 - \frac{TEMA(M)}{100}$ o Modelo de Descubrimiento de Reglas ▪ VP (C) = Número de elementos positivos del cluster C clasificados correctamente. ▪ VN (C) = Número de elementos negativos del cluster C clasificados correctamente. ▪ FP (C) = Número de elementos negativos del cluster C clasificados incorrectamente. ▪ FN (C) = Número de elementos positivos del cluster C clasificados incorrectamente. ▪ IE (C) = Índice de pertenencia de un elemento al cluster C. Mide la probabilidad de que un elemento pertenezca a un determinado cluster C. $IE(C) = \frac{VP(C)}{NR(C)}$

- TPR (C) = Número total de elementos positivos reales del cluster C. $TPR(C) = VP(C) + FN(C)$
- TNR (C) = Número total de elementos negativos reales del cluster C. $TNR(C) = FP(C) + VN(C)$
- Acierto: es la proporción del número total de casos predichos que son correctos. Mide el nivel de certeza del modelo.

$$\text{Acierto} = \frac{VP + VN}{TPR + TNR}$$

- SMC (M) = Sensibilidad del modelo de clasificación. Mide la proporción de casos positivos que fueron identificados correctamente, es decir la probabilidad que un elemento de una clase o cluster sea clasificado correctamente en esa misma.

$$SMC(M) = \frac{VP}{TPR} = \text{TasadeVerdaderosPositivos}$$

- ESMC (C) = Especificidad del modelo de clasificación. Mide la proporción de casos negativos que fueron identificados correctamente, es decir, la probabilidad que un elemento de una clase o cluster sea clasificado correctamente en la misma.

$$ESMC(C) = \frac{VN}{TNR} = \text{TasadeVerdaderosNegativos}$$

- PMC (M) = Precisión del modelo de clasificación. Mide la proporción de los casos predichos positivos que son correctos, es decir, la probabilidad de que una predicción efectivamente corresponda con su valor real. $PMC(M) = \frac{VP}{VP + FP}$
- EXMC (M): Exactitud del modelo de clasificación. Mide la proporción de casos correctamente predichos.

$$EXMC(M) = \frac{VP + VN}{TPR + TNR}$$

- Error ó Confusión: es la proporción de casos incorrectamente predichos. $\text{Error} = \frac{FP + FN}{TPR + TNR}$
- Tasa de Falsos Positivos y Negativos: es la proporción de casos incorrectamente predichos.

$$\text{TasadeFalsosPositivos} = \frac{FP}{TNR} \qquad \text{TasadeFalsosNegativos} = \frac{FN}{TPR}$$

- Medida F: es una medida estadística que combina las métricas de Precisión y Sensitividad para evaluar de forma más realista la certeza del modelo. $\text{MedidaF} = \frac{2}{\frac{1}{PMC(M)} + \frac{1}{SMC(M)}}$
- Coeficiente de Kappa: mide la precisión del modelo para predecir la clase o cluster verdadero. $k = \frac{P(\text{Acierto}) - P(\text{Error})}{1 - P(\text{Error})}$

  donde P (Acierto) es el porcentaje de aciertos y P (Error) es el porcentaje de casos incorrectamente clasificados.

o Modelo de Descubrimiento de Dependencias Significativas

En esta instancia de la investigación no se han encontrado métricas representativas para este modelo.

4.3. Métricas de Proyecto

Si bien en [Marbán, 2003] se definen las categorías que involucran las características más importantes de los proyectos de Explotación de Información, dentro de la cual se encuentra la categoría Proyecto, no se hace ninguna mención en ésta sobre los criterios de evaluación de los resultados obtenidos del proceso de minería de datos, para lograr un resultado exitoso del proyecto. Con lo cual, las métricas que se proponen para evaluar los resultados, se incluirán dentro de esta categorización. En la tarea de evaluación de los modelos, se estableció la métrica de exactitud como uno de los factores para evaluar la calidad de los modelos construidos. La exactitud de un modelo está relacionada con el grado de confiabilidad que tienen los resultados obtenidos frente al objetivo de minería de datos que se persigue y, en consecuencia, con el éxito del proceso de desarrollo del proyecto de explotación de información. Para definir la métrica que determina el éxito de un proyecto de explotación de información, se considerará el resultado de la métrica de exactitud para cada modelo M de minería de datos construido. A cada modelo M, se le asignará un peso según su nivel de exactitud, como se indica en la tabla 3:

Tabla 3. Pesos asociados a niveles de exactitud de modelos

NIVEL DE EXACTITUD	RANGO DE LA MÉTRICA	PESO
Muy Bajo	0 – 20%	0
Bajo	21 – 49%	0.25
Nominal	50 – 70%	0.5
Alto	71 – 90%	0.75
Muy Alto	> 90%	1

A partir del proceso de desarrollo de proyectos de Explotación de Información [Vanrell, 2011], se han considerado métricas de proyecto para las tareas que se indican en la tabla 4.

Tabla 4. Propuesta de Métricas de Proyecto

Subproceso: Evaluación
Tarea: Evaluar los resultados
Métricas Propuestas
▪ NM = Número de modelos construidos para el proyecto. ▪ NEM (M) = Nivel del exactitud del modelo M. Mide el nivel de exactitud de cada modelo construido para el proyecto de explotación de información. $NEM(M) = \frac{\sum Exactitud\_Modelos(M)}{NMDM(M)} * 100$ donde Exactitud_Modelos son los niveles de exactitud obtenidos para cada clasificación de modelo de explotación de información (descubrimiento de grupos, descubrimiento de reglas y descubrimiento de dependencias significativas). A cada modelo M, se le asigna un peso según la tabla 4. ▪ NMMA = Número de modelos clasificados como Muy Alto. ▪ NMA = Número de modelos clasificados como Alto. ▪ NMN = Número de modelos clasificados como Nominal. ▪ NMB = Número de modelos clasificados como Bajo. ▪ Éxito = Éxito del proceso de desarrollo. Mide el nivel de los resultados de minería de datos obtenidos respecto a los criterios de éxito del proyecto. Cuanto mayor sea el este valor, implicará una mayor aceptación por parte del usuario del proyecto de explotación de información. $Éxito = \frac{NMMA + (NMA * 0.75) + (NMN * 0.5) + (NMB * 0.25)}{NM} * 100$
Tarea: Revisar el proceso
Métricas Propuestas
▪ Si 0 <Éxito < 50% el proceso no responde a los criterios de éxito del problema de negocio. ▪ Si 51% <Éxito< 80% el proceso debe ser revisado y ajustado. ▪ Si 81 <Éxito< 100 el proceso cumple con los criterios de éxito del problema de negocio.
Subproceso: Entrega
Tarea: Producir un reporte final
Métricas Propuestas
▪ GDE = Grado de documentación a entregar. Mide el esfuerzo necesario para producir la documentación durante el proyecto. Los valores pueden ser: - Bajo: Sólo para los documentos implantados - Nominal: Sólo para modelos generados - Alto: Todos los modelos (generados e implantados) y subprocesos del proceso de desarrollo. - Muy Alto – Todos los modelos (generados e implantados) y procesos del proyecto de explotación de información (administración y desarrollo).

5. Conclusiones

Se definió una propuesta de métricas aplicable en proyectos de explotación de información para PyMEs, que permite evaluar su avance y calidad durante el proceso de desarrollo. Del trabajo realizado, surge que algunas tareas definidas en el proceso

de desarrollo [Vanrell, 2011] no permiten obtener métricas significativas que deban ser consideradas.

Se prevé tener una primer validación de las metricas propuestas en el marco de los proyectos de explotación de información que desarrollan los alumnos en la Asignatura "Tecnologías para Explotación de Información" en la Carrera de Licenciatura en Sistemas de la Universidad Nacional de Lanús y en la Carrera de Ingeniería en Sistemas de Información de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional.

Como futura línea de trabajo se estima continuar con la definición de métricas aplicables a la ponderación de atributos en el Modelo de Dependencias para Explotación de Información, y la validación de la solución propuesta en experiencias de trabajo tanto académicas como reales.

6. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A167 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina); y por el Programa de Incentivos a la Investigación del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza (Argentina).

7. Referencias

Britos, P. 2008. Procesos de Explotación de Información Basados en Sistemas Inteligentes. Tesis Doctoral en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata.

García Martínez, R., Servente, M. y Pasquini, D., 2003. Sistemas Inteligentes. Editorial Nueva Librería. Buenos Aires.

Hauge, O., Britos, P., García-Martínez, 2006. Conceptualization Maturity Metrics for Expert Systems. IFIP International Federation for Information Processing.

Marbán, O. 2003. Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCoMo). Tesis Doctoral. Facultad de Informática. Universidad Politécnica de Madrid.

Pollo Cataneo, M.F, 2007. Sistemas Expertos. Conceptualización y Métrica de Madurez. Trabajo Final de Especialidad en Ingeniería de Sistemas Expertos. ITBA Instituto Tecnológico de Buenos Aires.

Pytel, P., 2011. Método de Estimación de Esfuerzo para Proyectos de Explotación de Información. Herramienta para su Validación. Tesis de Magister en Ingeniería del Software. Universidad Politécnica de Madrid. ITBA Instituto Tecnológico de Buenos Aires.

Vanrell, J, 2011. Un Modelo de Procesos para Proyectos de Explotación de Información. Tesis de Magister en Ingeniería de Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.

Fractalizing Social Networks

Silvia Cobialca, Juan M Ale

Universidad Austral, Buenos Aires, Argentina
silvia.cobialca@accenture.com, ale@acm.org

Abstract. Fractals are self-similar structures that exist widely in nature. We are aiming the current work to prove that social networks, although not a naturally generated structure but one created by humans within the World Wide Web, show a fractal behavior as well and as such, will experience a self-similar kind of evolution.

In the present work we attempt to find through the study of fractal behavior, how the introduction of a new element in the social network will impact in the existing network structure and in the network growth. Also our main interest is into how the new node will start interacting with the existing communities in order to eventually build its own.

Keywords: Social Networks, Fractals, fractal dimension, box dimension, adjacency matrix, fractal social models and algorithms.

1 Introduction

The main focus of this work is the idea of social networks as fractals and how to explore and elaborate from there in order to build new functionality useful to the study of their behavior over time. If social networks behave like fractals then, the existing network random generators would not apply to them as their randomness lose their meaning and we will need specific generators with proper parameters more suitable to be applied to fractal behavior.

We show how fractal nature applies to social network structures and how their evolution can eventually be predicted by modeling upon them. To do so, our starting point is to reviewthe basics about networks in general and their parameters, measures, types and the existing models used to generate network structures, to later focus on social networks in particular. We also review fractal theory and its applications to be able to merge the two concepts together while working on the models.

As a motivation, we noticed that current models for networks are based on randomness in a general way and sometimes such models donot take into account the nature of the network. In that sense, we consider that social interaction has a strong relevance that should be taken into account when studying models that will be used in the future for social networks. We base the present work on the basic structure of the social network and escalate from there in order to formulate the prediction as precisely as possible thinking more in the individual components and how their behavior will imprint its pattern in the whole network.

Our main contribution is a model and algorithms that allow us to show that evolution of a social network obeys fractal rules.

After reviewing previous work on this matter we have come with the ones more relevant to our line of thought which are detailed below:

Leskovec (2008) studied network evolution, network cascades and large data while analyzing large social networks as a whole in order to formalize and try to predict future behavior and structure. His idea was very nicely worked out by a three by three focus on observations, models and algorithms on network evolution, cascades and large data. He worked mostly on the network itself, but not from an individual node point of view.

Song et al. (2005) analyzed several real networks and found that they consist of self-repeating patterns on all length scales.

In the work by Faloustos et al. (2006),the authors also investigated network structures and expanded about the graph generators, but although they mentioned power laws they didn't elaborate on the possibility of using fractal similarity to build the network synergy expressed by the links between nodes.

Erdös, P., Renyi, A.(1959) set the base for future contributors to the area by defining random network generation and probability of connections between nodes among other definitions.

We use a different approach from the previous work mentioned above as our main focus is the idea brought by Benoit Mandelbrot (1982). The author defined fractal structures and the fact that they were present in nature in several levels of complexity; he also mentioned that the interaction between systems can be seen as fractals. This is, in fact, the driver of our present work.

The remainder of this article is organized as follows. In Section2 we provide background information on the main topics of this work. In Section 3 we present our Fractal Social Network Model and our proposed algorithms. Some empirical results are shown and discussed in Section 4. Finally, we conclude our work in Section 5

2 Background information

Social networks.

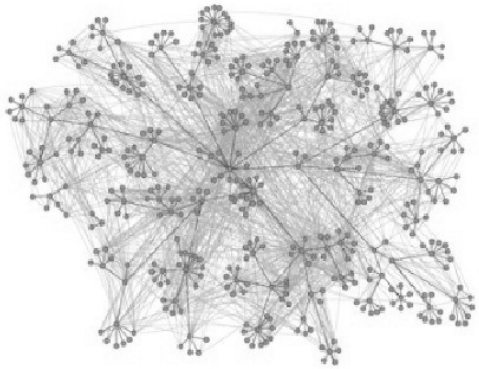
Actual social networks are more related to collections of social ties among friends, like the ones based on the internet as Facebook, Twitter, Instagram, Tumblr, Flickr and many others; although there are other networks focused on businesses or professional relationships like LinkedIn, WordPress, Yelp, etc. These examples are all among online networks which, because of their nature, can be massive and reach farther boundaries than those based on people direct interactions, which can be geographically based, although both kinds are good examples of social networks where the same theory can be applied.

Social interaction have grown steadily in complexity over the course of human history, due to technologicaladvances facilitating distant travel, global communication, and digital interaction as mentioned by Kadushin (2012).

These networks can be seen as graphs with nodes (the individuals participating in the network) connected by links as shown in Figure 1.

Understanding any one piece of information in this environment depends on understandingthe way it is endorsed by, and refers to, other pieces of information within a large network oflinks.

Fig.1.*A representation of a social network based on email communications (Image from http://www.personal.umich.edu/ladamic/img/hplabsemailhierarchy.jpg)*



We represent social networks as graphs because a graph is a way of specifying relationships among a collectionof items. A graph consists of a set of objects, called nodes, with certain pairs of theseobjects connected by links called edges.

Graphs are defined as *directed* if they consist of a set of nodes together with a set of *directed edges* each directed edge is a link from one node to another, with the direction being important. Directed graphs are generally drawn as in with edges represented by arrows. When we want to emphasize that a graph is not directed, we can refer to it as an *undirected* graph.

Random graph algorithms

From the initial work done by Erdös and Renyi(1959) to the present, several algorithms have been developed in order to generate social network graphs (or graphs in general) and study their evolution over different epochs (a way to call the parameter to measure passing time). The simplest algorithm was one of complete randomness where the probability of two nodes connecting (or contacting) each other was the same for every pair of nodes belonging to the network.

Of course, this approach is too simple and no real life network will behave that way, more likely the nodes involved within the network will connect with other nodes based on preferences, similarities, recommendations from others but they hardly will connect in a random fashion.

There are also other kinds of social networks like the ones based on*small-world phenomenon* in which the applied logic states that any two individuals in the network are likely to be connected through a short sequence of intermediate acquaintances. This has been proved to be truth by several previous investigations being the oneconducted by Milgram(1967) the most popular, but we also have reviewed the work from Mathias and Gopal (2000), as theyall reveal that often we meet a stranger and discover that we have an acquaintance in common. Recent work has suggested

that the phenomenon is also existent in networks arising in nature and technology, and a fundamental ingredient in the structural evolution of the World Wide Web. We explore if the "fractal" network as we call it, exhibits also a sort of small-world phenomenon in its behavior.

What are fractals?

First of all, and before getting to the point of a definition of a fractal let's take a moment to imagine what we currently denote as chaos, or chaotic behaviorusually related to some unpredicted pattern that cannot be formalized in any way,as studied by Shroeder (1991),. The difficulty of working with such behavior is, of course, the inability to adjust it to any existent law that could rule it and help to the job at hand.

Fractals are self-similarity structures that can explain this behavior and bring certainty and predictability whereas there was chaos and misinterpretation beforeThe trick is to understand the structure itself and how it evolves and grows from the basic initial unit.

Fractals are useful in modeling and explaining natural complex patterns that can't be explained by Euclidean geometry. In these irregular and fragmented patterns, we can see how nature expresses itself in leaves, mountains, turbulences and also inside of us in our blood vessels or pulmonary systems. They are all examples of shapes that can be built by scaling up a base structure over and over, which implies a certain degree of irregularity, but in an unusual regular way at all scales.Barnsley (1988) in his work compiled different fractals existent in nature like forests, mountains and landscapes in different parts of the world, and also reviewed the theory behind their existence.

To understand that, we need to find the fractal dimension and the basic structure that the fractal is built upon and later grow it from there.

Defining the fractal dimension.

Fractal Dimension.

The concept of fractals started to take form when Benoit Mandelbrot vocalized his idea of a continuous escalating structure found in nature in different organic and inorganic systems.

The fractal dimension is a measure that will indicate the relationship between the size of the individual smallest structure that comprises the fractal and the total size. The formula is as:

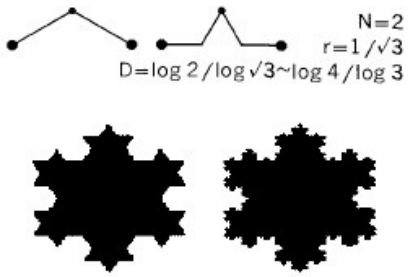
$$D=\log N(r)/\log (1/r) \quad (1)$$

Where
D: Fractal (or Hausdorff) Dimension
N: number of base parts
r: similarity ratio

There are a few examples that can help us illustrate this definition in a way that clearly enlightens our knowledge and understanding of it. One example is the Koch

coastline shown in Figure 2.Every line the triangle is made of is added another pair of lines in the middle and the pattern is repeated over and over in every side finally getting the Koch triangle in black below the main structure.

Fig.2. Triad Koch coastline Fractal Dimension



Box Dimension.

There is another way to measure the fractal dimension and it is by sizing the smallest squared box that will include the base structure and counting the amount of those boxes that fit in the whole fractal structure. The number of boxes is N(r) where r is the size of the side of the square used as box dimension. Then we can calculate the fractal dimension using (1).

Multi-fractal Dimension.

As with the normal fractals defined above, there exists also a category of fractals that donot have one only fractal dimension and because of their nature, since they were built from, say, bricks of different sizes, they can be called multi-fractals.

Hence, the trick with multi-fractals is to identify the base "bricks" they are made of. Examples of multi-fractal structures are the diffusion-limited aggregations (DLA) like the ones generated by the colloids where the structure grows one molecule at a time. An image of a multi-fractal is shown below in Figure 3.

Fig.3. An example showing a multi-fractal

Are all social networks fractals?

After reviewing the information provided above, we can start to think that social networks seem to behave like multi-fractal structures. We cannot at this point of our investigation ascertain which would be the fractal dimension of them and more experimentation is needed for us to be able to provide such metrics.

For now, let's just hypothesize about how the behavior and the way they get generated seem to be compliant with that of multi-fractals.

3 The Fractal social network model

3.1 The model.

First of all, it has been observed during the initial investigations that the introduction of a new node v in the network is always through the knowledge of at least one pre-existent node w and as that happens, even if the new node has been introduced into the whole network, only a relative small region is available to it. Nobody expects the newcomer to start interacting with every community in the network right away, and for that interaction to start, time is of essence. We call that amount of time $T_0$the "introduction time" which is a grace period allowed to the new node before it gets the first contact with other nodes existent in a network and that is different than its sponsor in the immediate network (or community). The introduction time depends on the size of the community where node w exists, and of course the ties between v and w which are unknown and different to every case, so this is another parameter to take into account at the time of the on-boarding of the new node.

Once node v starts its interaction with the nodes in w's community, there will be only a certain amount of time T_1 in which it will start to reach out to nodes belonging to the neighborhood in a way that will be proportional to the amount of connections its newly acquired community has with the outside communities. This means that, if node w community C_0 has connections with three other communities C_1, C_2 and C_3 but the connectedness between them is, say n_1, n_2 and n_3 in which $n_1 >> n_2 >> n_3$ then the interactions between v and the nodes in said communities will start before with C_1 than with C_2 or C_3.

Now, the next item to take into consideration is the kind of node v will become within time, meaning if it will be a popular node or the opposite, more like a shy node. That will depend on the willingness it has to share and activate new connections with the rest of the network. In other words, if v is already a popular and well known individual when it gets in the network, there is a great possibility that it will "attract" the attention of the other nodes and they will reach out to it to connect and become popular as well. But if v is an individual that got in the network for a specific task and nothing else, it is improbable that its connections will increase beyond what is expected from it and so its degree of connectedness will be very small.

Fromthe adjacency matrix and the fractal dimension we get the minimum structure to be replicated in order to get what becomes the node evolution and future

participation in the network. This can be done in several iterations and with different nodes in order to get the final base structures of the network.

We are talking about base structures because we consider the social network to be a multi-fractal and as such, it is built upon several box dimensions. We describe in detail this procedure in the next subsection.

There is a pre-condition to be taken into account in this algorithm, which is that a network should exist prior to the application of this algorithm. This way, the new incoming nodes can have the base layers of the existent network and they canreplicate similar pre-existent structures as theyactivate their new connections.This is key to ensure self-similarity patterns.

3.2 The fractal connection algorithms

We present the GetBox algorithm which will get the box dimension specific to a node in a certain community. It finds all the nodes the given node is likely to connect in order to keep the self-similarity structure in the network and its current connections.

Algorithm: *GetBox(N,Adj[N,N],v,Cm[N])*

```
Returns Bv list with nodes for potential connections

For each v in Adj[N,N]

Oudg[v]:=outdegree_t(v,Cm(v))--the outdegree of v within its
communities of interest

End For

Freq:=Frequency(Oudg(v))   --we calculate the frequency
distribution of all nodes' outdegreeconnections

Freq=Rnd(Fr)      --we select randomly one of the existent
Frequencies in the network to be the box dimension for i at t+dt

Create Empty List(Bv)

i=v

for each j in Adj_t[N,N]

If Adj_t[i,j]=1or j in Cm[i] and Cardinality(Bv)<Freq--the
frequency is a measure of the box dimension, since there are several
boxes we use one of the available options randomly

Then Add j to B_v     --B_vwill contain all the nodes v should be
connected to in t+dt

End For

End.
```

The ApplyBox algorithm will use the Box dimension obtained by the GetBox algorithm in order to apply those new connections to the given node (and keep the existent ones)

Algorithm: *ApplyBox(v,*Adj_t*[N,N],*w_f*,*B_v*)*

```
We assume v has been in the network since time t
Begin
i:=v        --we search in row corresponding to v node
For each jin Adj(t+dt)[N,N]
If Adj(t+dt)[i,j]=0
ifj in Bv and wf(j)>0.6  --the higher the wf the more
willing to connect and the longer will be Bv list
Adj(t+dt)[i,j]=1
End For
End.
```

List of variables

Willingness Factor: w_f (used to distinguish nodes with interest of acquiring new connections from others not that interested in any interaction)

Size of the network: N (number of nodes in the network)

Introduction time: t (the time at which the new node is included in the network)

Adjacency matrix: Adj [N,N] (a matrix which describes a graph by representing which vertices are adjacent to which other vertices)

Cm array: community [N] (we don't know at first how many communities exist, at a maximum it can be the same as the amount of nodes)

Box dimension list: B_v(list with nodes that will likely get connected to the new node in order to preserve self-similarity in the network)

4 Experimentation: A simple example and findings

To experiment our theory we have a Facebook network of 484 nodes and 33272 links where we are able to see interactions between two timeframes.

We collected the network and classified the nodes with more activity, the new nodes and the nodes that left the network between timeframes *t* and *t+dt.* The network is shown below in figure 4.

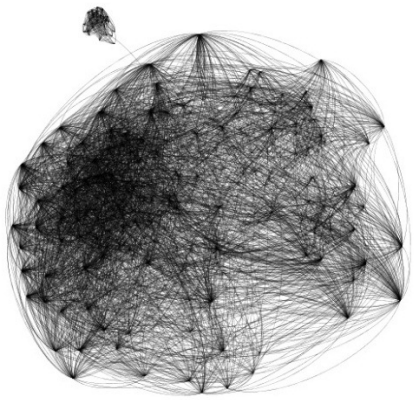
Below are listed the key findings during our experimentation:

- The willingness is a very important parameter to be considered in the existent nodes as well as in the newcomers. It has relevance in the connections growth because if the nodes are not interested in connecting with other existent nodes

they will not attempt any new interactions, even if they have the best opportunities by being linked to the most popular nodes and communities.

- The frequency distribution of outbound connections (we are for now interested in the outbound) seems to be related to the different fractal dimensions of the boxes.
- There were several nodes observed at t with sponsors among the nodes with more connectedness and they still didn't increase their degree of connection at $t+dt$. We believe that these nodes weren't interested in new connections.
- One of the nodes with highest degree of connections didn't increase its connectedness in the new timeframe which contradicts the "rich get richer" principle. This behavior can also be explained by the willingness factor in a way of saying that there were no interesting things out there for this node to grow to.
- There are also some nodes that shrunk during the timeframe considered in our study, which is not a surprise as due to the network dynamics it's expected a certain level of change in both ways for the nodes degree of connections. This fact is still worth mentioned and something to be investigated and expanded in the future as well.
- The nodes with more growth grew beyond their community of origin and we can infer that they are then more mature in a way that they can reach and interact with new communities. It is expected that in the future they will continue to grow in this same way. The new communities were known from the most popular communities in the list of the pre-existent linked nodes. This is exactly what we are taking into consideration while developing theGetBox and ApplyBox algorithms.

Fig. 4. The network used in the experimentation phase (notice the new small community at the top left of the graph)



5 Conclusions and future work

This is a work in progress and in the following paragraphs we are presenting the first results of this research.

We have come to understand from our work during the experimentation phase and the theoretical background that some of the interactions in social networks can't be taken as random and more so, the people making the present social networks act some times in ways that seem to be mimicking other people behavior. Hence our self-similarity approach seems to be more suitable for them than randomness.

We have also uncovered the existence of several parameters to be taken into account when modeling social network.

We are leaving for future work and enhancements of the model, the task of testingthe algorithms presented in this work in a more global social network, also the formalization of rules to prevent starvation of network components in order to ensure that all the components are added to the fractal structure while the evolution happens as well as how we explain the dynamics of the shrinking patterns in the nodes connections.

6 References

1.Leskovec J.: Doctoral Thesis: Dynamics of large networks. CMU-ML-08-111 (2008)

2. FaloustosC.,ChakrabartiD.: Graph Mining: Laws, Generators and Tools. ACM Computing Surveys, vol 38 (2006)

3.Shroeder M.: Fractals, chaos and power laws. W. H. Freeman and Company (1991)

4. Mandelbrot, B.: The Fractal Geometry of Nature. International Machine Business (1977)

5. Barnsley M.: Fractals Everywhere. School of Mathematics. Georgia Institute of Technology (1988)

6. Mathias N., Gopal V.: Small Worlds: How and why. Department of Computer Science and Automation (2000)

7. Erdös, P., Renyi, A.: Of the Evolution of Random Graphs. (1959)

8. MilgramS., The small world problem. Psychology Today (1967)

9. Burt R.: Structural Holes and Good Ideas (2004)

10.Kadushin C.: Understanding Social Networks. Oxford University Press (2012)

11. Song, C., Havlin, S., Makae, H.: Self-similarity of Complex Networks. Nature, Vol 433. Nature Publishing Group (2005).

Determinación de género y edad en blogs en español mediante enfoques basados en perfil

Dario G. Funez, Leticia C. Cagnina y Marcelo L. Errecalde

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Facultad de Ciencias Físico, Matemáticas y Naturales,
Universidad Nacional de San Luis - Ejército de los Andes 950
(D5700HHW) - San Luis - Argentina Tel: (0266) 4420823 / Fax: (0266) 4430224
e-mail: {dgfunez, lcagnina, merreca}@unsl.edu.ar

Resumen La determinación del perfil del autor (desconocido) de un documento permite identificar características como el género (sexo) y edad de dicho autor, en base al estilo de escritura y las palabras presentes en el documento. Esta tarea está creciendo en importancia en diferentes áreas de investigación, especialmente en idiomas como el español donde existen pocos trabajos realizados hasta el presente. En este trabajo se presentan los resultados obtenidos en la determinación de género y edad en blogs en español mediante *enfoques basados en perfil*, un tipo de técnica que ha sido aplicado exitosamente en tareas de atribución de autoría. Los resultados obtenidos muestran la viabilidad de la aplicación de este enfoque en la determinación del perfil del autor de un documento y permiten identificar aspectos que necesitan ser mejorados en el futuro.

Palabras Claves: categorización de documentos, minería de textos, perfiles de autores, enfoques basados en perfiles

1. Introducción

El uso creciente de redes sociales como `Facebook` y `MySpace`, sitios de microblogging como `Twitter` y las innumerables facilidades de chats disponibles hoy en día han hecho accesible mucha información provista por personas de diferentes edades, género, condición social, etc. Dicha información puede ser utilizada para inferir datos importantes del perfil del autor de un texto como su personalidad, demografía y antecedentes culturales [2]. La determinación del perfil del autor de un documento (en inglés *author profiling*) es la tarea de distinguir entre clases de autores en base a la forma del lenguaje compartido por un grupo social particular. Esto puede involucrar la identificación de diversos aspectos del perfil de una persona tales como el *género* (femenino vs masculino), *edad* (de acuerdo a distintos grupos etarios), *lenguaje nativo* y *tipo de personalidad*.

La actividad de recabar información del perfil del autor de un documento es un problema de interés creciente en áreas como seguridad y anti-terrorismo, marketing y diversas disciplinas forenses. En el caso de marketing, es evidente

el beneficio que puede obtenerse del empleo de los comentarios de clientes en blogs para determinar la demografía de la gente que gusta o no de determinados productos [1]. Asimismo, este tipo de técnicas también pueden tener un impacto importante en problemas forenses de abordaje más reciente como la detección automática de depredadores sexuales en la Web [7].

Por otra parte, en problemas de *atribución de autoría* [10] han ganado cada vez más relevancia los enfoques *basados en perfiles* [3,4,5,6] planteándose como alternativas interesantes a los enfoques clásicos de categorización de textos *basados en instancias* [10] debido a diversas ventajas como su facilidad de implementación, eficiencia de aplicación, escalabilidad y la representación explícita de información relevante sobre un autor.

En este trabajo analizamos si los enfoques basados en perfiles son adecuados para la determinación de la edad y el género de documentos de blogs en idioma español. Con respecto a la edad, se consideran posts de 3 grupos etarios distintos formados por adolescentes entre 13 y 17 años, jóvenes entre 23 y 27 años y adultos entre 33 y 47 años. El género por su parte puede ser femenino o masculino.

Estos enfoques fueron probados con la versión en español del corpus de entrenamiento del *PAN-PC-2013* [1] por ser el único disponible para experimentación a la fecha. Los resultados obtenidos sustentan la viabilidad de este tipo de técnicas y son un punto de comienzo para mejorar su desempeño en trabajos futuros.

El resto de este trabajo se organiza de la siguiente manera. En la Sección 2 se explican los principales conceptos vinculados a los enfoques basados en perfiles. La Sección 3 detalla los principales aspectos de la tarea abordada y cómo los enfoques basados en perfiles se ajustan a su resolución. La Sección 4 describe el trabajo experimental y el análisis de los resultados obtenidos. Finalmente, en la Sección 5 se exponen las conclusiones de nuestro estudio y se proponen trabajos futuros para mejorar esta propuesta.

2. Enfoques basados en perfiles

Los *enfoques basados en perfiles* constituyen uno de los enfoques principales hoy en día para abordar problemas de *atribución de autoría* (AA) [10]. En un problema de AA típico, un texto de autoría desconocida es asignado a un autor candidato, dado un conjunto de autores candidatos para los cuales se tienen disponibles textos de muestra de autoría indiscutida. Desde un punto de vista del aprendizaje automático (*machine learning*), esto puede ser visualizado como una tarea de categorización de texto de múltiples clases y único rótulo (*multi-class single-label*). En este contexto, para cada clase (autor) se construye un perfil de autor que contiene información recuperada de un conjunto de documentos escritos por el mismo [4]. En la la parte izquierda de la Figura 1 se muestra gráficamente el proceso de generación de los perfiles de cada autor.

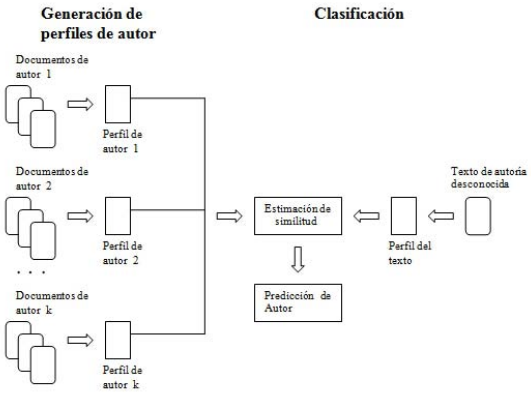


Figura 1: Atribución de autoría basada en perfil: generación de perfiles de autor (izquierda) y clasificación de un documento de prueba (derecha).

Las características a recuperar de un texto pueden estar basadas en el *estilo* de escritura o en el *contenido* del texto.

- *Características basadas en estilo*: extraen de los documentos medidas estilográficas como la frecuencia de determinadas clases de palabras como pronombres, artículos, preposiciones, cantidad de hipervínculos, promedio de palabras en un post etc. [9]. Estas características usualmente varían dependiendo del género y la edad. Por ejemplo, las mujeres en los blogs suelen utilizar más pronombres y palabras de aprobación y negación, reduciéndose esta tendencia en edades más avanzadas.
  Una de las características más utilizadas para capturar aspectos de estilo son las frecuencias de n-gramas de caracteres. Los n-gramas son subcadenas de n caracteres consecutivos [3] siendo común el uso de $n = 3$, 4 y 5. En el inglés por ejemplo, el uso de tri-gramas de caracteres demostró ser efectivo para capturar la frecuencia de adverbios, información contextual, etc.
- *Características basadas en contenido*: consideran las palabras que pertenecen a temáticas particulares [9]. En esta categoría también existen diferencias en su uso dependiendo del género del autor. Por ejemplo, las escritoras tienden a utilizar con más frecuencia palabras relacionadas con lo personal como por ejemplo “shopping”, “madre”, etc. En cambio, los escritores se interesan más en temas relacionados a política y tecnología. Los adolescentes escriben sobre sus amigos, estados de humor y temas relacionados al colegio. A edades mayores crece su interés por el casamiento, la política y temas financieros.

Para obtener el perfil de un autor se considera un conjunto de documentos de su autoría y se extraen de él un conjunto de L características. Por ejemplo, para el caso en el que se seleccione como característica los tri-gramas (subcadenas de 3 caracteres), se deben obtener todos los tri-gramas de cada documento y se los ordena por su frecuencia. Luego, los L tri-gramas más frecuentes constituirán el perfil. Para poder clasificar un documento a un autor, se necesita generar el

perfil del documento y luego, utilizando una medida de distancia (o de similitud), se determina si existen similitudes entre el perfil del documento y el perfil de cada autor [6]. El autor cuyo perfil sea el más "cercano" al perfil del documento, será el que se retorne en el proceso de clasificación, como se muestra en la parte derecha de la Figura 1. Algunas de las medidas de distancia/similitud utilizadas en los enfoques basados en perfil son las siguientes:

- *Keselj's Relative Distance (KRD)*: esta medida, referenciada en algunos trabajos como *CNG*, mide la distancia K entre dos perfiles P_1 y P_2 como

$$K = \sum_{x \in X_{P_1} \cup X_{P_2}} \left(\frac{2 \times (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2$$

  donde $P_i(x)$ es la frecuencia del término x en el perfil P_i, y X_{P_i} es el conjunto de todos los términos que ocurren en el perfil P_i.
- *Simplified Profile Intersection (SPI)*: Esta medida de similitud es una versión simplificada de la anterior, que sólo toma en cuenta la cantidad de características que pertenecen a ambos perfiles [5].
- *Out of Place (OOP)*: Estima la diferencia posicional de cada característica en los perfiles a comparar y la medida es la suma de todas las diferencias posicionales para todas las características de los perfiles [3].

Un inconveniente con las medidas enunciadas previamente es que muchos de los términos que ocurren en los perfiles de los autores son términos utilizados muy frecuentemente en el lenguaje apareciendo en consecuencia en todos los perfiles y por lo tanto aportando poca información discriminativa entre un perfil y otro. Esta observación, ha conducido a nuevas propuestas como el *recentrado de perfiles locales* que se han aplicado recientemente con éxito en tareas de AA [6] y consideramos que pueden ser relevantes en el problema abordado en el presente trabajo, por lo que será descripto en forma más detallada a continuación.

El *recentrado de perfiles locales* (RPL) crea perfiles priorizando aquellas características que son usadas en forma diferencial respecto al uso del lenguaje habitual [6]. Para medir el uso del lenguaje habitual, el RPL utiliza un *perfil del lenguaje* que aproxima el uso real de las características en el lenguaje midiendo las frecuencias de ocurrencia en todos los documentos del conjunto de entrenamiento. Si bien ésta es sólo una de las posibles implementaciones alternativas para usar como perfil del lenguaje, en [6] ha demostrado ser bastante efectiva.

De esta forma, para construir el perfil de un autor en RPL se utilizan las L primeras características, ordenadas (en forma decreciente) por el valor absoluto de la diferencia entre su valor de frecuencia de uso en el perfil del autor y su frecuencia de uso en el perfil del lenguaje.

El algoritmo RPL no sólo realiza un "recentrado" de los valores de los perfiles de cada autor respecto al perfil del lenguaje; también utiliza una función de distancia específica que se define como:

$$d(f_1, f_2) = \sum_{x \in profiles} \frac{(f_1(x) - E(x)) \times (f_2(x) - E(x))}{\|f_1(x) - E(x)\| \times \|f_2(x) - E(x)\|}$$

donde f_1 y f_2 son los perfiles a ser comparados, $f_i(x)$ es la frecuencia normalizada de la característica x en el perfil f_i, E es el modelo del lenguaje y el término $profiles$ denota el conjunto de las características que se encuentran en las primeras L posiciones de f_1 o f_2. Como se puede observar, esta medida no es más que la distancia coseno de las cuentas recentradas de ambos perfiles. Esto implica que el perfil del documento a ser clasificado también deba ser recentrado bajo este esquema.

3. Descripción del sistema clasificador

Nuestro estudio en este trabajo se enfoca en analizar la viabilidad del uso de enfoques basados en perfil, los cuales se han desempeñado con gran éxito en tareas de AA, en tareas de determinación del perfil de un autor de documentos en español. Específicamente, esta tarea consistirá en predecir el género del autor (femenino o masculino) y la edad del mismo (grupo de los 10's, 20's o 30's). El grupo de las edades de "10" comprende las edades entre 13-17 años, el de 20 entre 23-27 y el de 30 entre 33-47 años [1].

Como se puede observar, un perfil para cada una de estas clases no representará un autor particular sino una *clase de autores* de acuerdo a su grupo etario o el sexo de la persona. Así, por ejemplo, un perfil que se construya con documentos rotulados como "20" representará características de autores cuya edad oscila entre los 23 y 27 años en lugar de características de un autor particular.

Para usar un enfoque basado en perfil para determinar el género y la edad de un autor, se deben definir dos aspectos principales. En primer lugar, si los perfiles que se utilizarán consideran a cada una de estas subtareas (determinar el sexo y la edad) como dos tareas separadas o no. En el primer caso, se definirán perfiles separados para la determinación del género (uno para femenino y otro para masculino) y para la determinación de la edad (un perfil por cada grupo etario, tres perfiles en total). En el segundo enfoque, se considera que el hecho de tratar la edad y el género en forma conjunta y simultánea puede ser beneficioso y se definirá un perfil para cada una de las 6 posibles combinaciones de las categorías consideradas: "femenino-10", "masculino-10", "femenino-20", etc.

El otro aspecto a definir es el tipo de característica a utilizarse en los perfiles: palabras completas, n-gramas de caracteres, características estilográficas, etc. Los experimentos preliminares mostraron que el uso de palabras completas en lugar de n-gramas de caracteres y la utilización de perfiles de categorías combinadas ("género-edad") producen mejores resultados por lo que se describirá en el resto de esta sección cómo se implementó este enfoque.

El sistema completo se implementó en dos etapas. En la primera se generaron los perfiles para cada una de las 6 categorías combinadas consideradas: *masc-10, fem-10, masc-20, fem-20, masc-30* y *fem-30*. Para implementar el enfoque recentrado se debió obtener además el perfil del lenguaje. La característica que se utilizó para generar los perfiles es la de palabras completas, la cual demostró ser superior a distintas variantes de n-gramas de caracteres como las de 3-gramas, 4-gramas, 5-gramas y combinaciones de ellas (3-gramas y 4-gramas, 4-gramas y

5-gramas). Los documentos empleados para construir los perfiles, forman parte del corpus español de entrenamiento de la competencia *PAN-PC-2013* [1].

El perfil del lenguaje se obtuvo realizando los siguientes pasos en secuencia considerando todo el conjunto de entrenamiento:

1. *Generación de perfil por cada documento*: Para cada documento del conjunto de entrenamiento se obtiene su perfil de palabras, donde cada palabra tiene como valor asociado su frecuencia de ocurrencia en el documento. Estos perfiles fueron generados con la librería Morphadorner[1].
2. *Unificación de los perfiles en un único perfil*: En esta tarea se concatenan todos los perfiles obtenidos en el paso anterior, obteniéndose un perfil que repite una característica si ésta ya aparece en otro documento.
3. *Eliminación de entradas repetidas*: El perfil del lenguaje se logra sustituyendo las entradas de las palabras repetidas en una única entrada, con un valor que es el total de sumar todos los valores de esas entradas, normalizados por el número de documentos. De esta manera se calcula la acumulación de todos los valores de una característica.

Para obtener el perfil de una categoría se aplican las mismas tareas que para obtener el perfil del lenguaje, pero restringiéndose a los documentos propios de esa categoría. Si se trabaja con perfiles recentrados, se debe recentrar el valor de cada característica con los valores del perfil del lenguaje, y posteriormente se ordena considerando el valor absoluto del valor recentrado.

La segunda etapa del sistema clasificador consiste básicamente en la implementación del proceso de clasificación de un documento de test arbitrario d_t utilizando los perfiles obtenidos en la primera etapa. El clasificador recibe como parámetro de entrada un archivo *xml* que tiene el formato de una conversación en un blog. Este archivo, en los enfoques basados en perfiles sin recentrado recibe un preprocesamiento básico y se genera su perfil de documento, clasificándose de acuerdo al esquema mostrado en la parte derecha de la Figura 1, usando el perfil de categoría más cercano de acuerdo a alguna de las funciones de similitud/distancia descriptas en la Sección 2 (KRD, SPI u OOP).

En un método como RPL que utiliza recentrado de perfiles, el procedimiento de clasificación es un poco más complejo como se muestra en la Figura 2 y se describe a continuación:

- *Preprocesamiento del archivo de entrada*: Se eliminan los tags de las conversaciones y se sustituyen las imágenes por tres caracteres (IMG) para no perder información sobre el contenido en el documento. También los dígitos fueron reemplazados por un caracter especial definiendo patrones a reemplazar en el documento.
- *Generación del perfil del documento ordenado por frecuencias*: En este módulo se consigue el perfil de palabras del documento a clasificar.
- *Recentrado respecto del perfil del lenguaje*: En esta tarea se recentra el valor de todas las palabras del perfil del documento.

[1] *Morphadorner* es una librería escrita en lenguaje Java, de acceso libre para PLN y suministrada por la Universidad de Northwestern.

- *Chequeo por similitud (o disimilitud) con los perfiles de cada categoría*: Se compara el perfil del documento con el de cada categoría, retornándose el rótulo de aquella que es más cercana (menos distante).

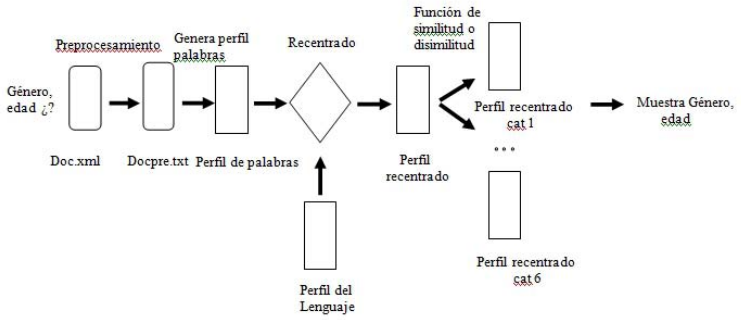


Figura 2: Diagrama del sistema clasificador utizando recentrado.

4. Experimentos

Para evaluar el desempeño de los distintos enfoques de clasificación basados en perfiles se obtuvieron dos conjuntos de prueba extraídos del corpus de entrenamiento de la competencia *PAN-PC-2013* [1]. El primer conjunto se utilizó para evaluar el género y está compuesto de 1000 documentos, con 500 archivos femeninos y 500 masculinos. En cada categoría, 166 archivos son de tamaño pequeño, 166 medianos y 166 grandes (considerando el tamaño en bytes del documento). El criterio utilizado para generar un conjunto de prueba con estas características, fue el de tener una colección con un número balanceado de documentos representativos de las distintas categorías, tamaños de documentos, etc. El otro conjunto de prueba se utilizó para evaluar la edad con un total de 1000 archivos divididos en tres categorías (10s, 20s, 30s) de 333 archivos aproximadamente cada uno, cada categoría también balanceada en el número de documentos de distintos tamaño.

Para que el clasificador sea aceptable, debe superar el "baseline" de 0.50 para el género y 0.33 para la edad, cuyos valores corresponden a una selección equiprobable entre las categorías disponibles y que también ha sido utilizado en la competencia. Los enfoques basados en perfiles que consideraremos en nuestro trabajo son los que construyen los perfiles en forma "clásica" tomando los L términos (palabras completas) más frecuentes. Utilizaremos en este caso, como medidas de similitud/distancia las siguientes 3 medidas que se describieron en la Sección 2: KRD, SPI y OOP. También analizaremos el desempeño de un enfoque que trabaja con perfiles recentrados como es el caso del algoritmo RPL. Respecto a este último punto, una pregunta que surgió al realizar este estudio, fue cual sería el impacto de trabajar con un perfil recentrado (como en RPL) pero en lugar de utilizar la función de distancia (coseno) propia de RPL medir la similitud/distancia entre los perfiles utilizando algunas de las funciones que utilizamos previamente en los enfoques no recentrados (SPI u OOP). Al uso de estas

funciones de distancia combinado con el recentrado del perfil los denotaremos SPI^{re} y OOP^{re} respectivamente.

En la Tabla 1 se muestran los resultados de la experimentación para el género en español, tomando como referencia el porcentaje de aciertos (clasificaciones correctas, en inglés *accuracy*) obtenido con los distintos enfoques basados en perfiles. Para cada uno de ellos, se especifica el resultado obtenido con distintos valores de L (tamaño del perfil) desde 200 hasta 8000. Como se puede observar, un enfoque basado en recentrado como RPL no obtiene resultados superiores al *baseline* (0.5) para ninguno de los valores de L considerados. Algo similar se observa cuando se usan perfiles recentrados con OOP como función de distancia (enfoque OOP^{re}). En este sentido, el único caso en que el uso de perfiles recentrados obtiene resultados por encima del baseline es el de SPI^{re}.

Respecto a los enfoques que utilizan los perfiles "clásicos" (KRD, SPI y OOP), estos obtienen mejores resultados superando al baseline con todos los valores de L considerados. En algunos casos incluso, se obtienen valores cercanos o levemente superiores a 0.6 que son comparables a valores preliminares reportados en tareas similares de categorización de género en blogs [1]. Es importante remarcar que con algunas medidas de distancia como SPI, el incremento progresivo del valor de L por encima del tamaño máximo reportado en este trabajo (8000), puede generar mejores porcentajes de aciertos con el conjunto de prueba considerado. Así por ejemplo, usando SPI con $L = 40000$ se logra una "accuracy" de 0.673. Se debe tener en cuenta sin embargo, que esto se logra a costa de un incremento significativo del tiempo de CPU requerido en tiempo de prueba, los cuales pueden ser inaceptables cuando se deben categorizar decenas de miles de documentos. Por otra parte, este incremento del L para mejorar la accuracy, también podría generar un efecto de "sobreajuste" (en inglés *overfitting*) a las características particulares del conjunto de prueba utilizado. Elegir un valor de L adecuado para lograr un balance correcto entre tiempos de clasificación aceptables, buen porcentaje de aciertos y evitar el efecto de sobreajuste es un factor importante que será abordado en trabajos futuros.

Respecto a la determinación de la edad del autor en blogs en español, los resultados de la Tabla 2 confirman lo observado previamente respecto a la baja efectividad del uso de perfiles recentrados en este tipo de tareas. En los 3 casos que utilizan perfiles recentrados (RPL, SPI^{re} y OOP^{re}) ninguno de los valores de L utilizados permitió superar el baseline de 0.33 para esta tarea.

También aquí, KRD, SPI y OOP superan ampliamente el baseline para esta tarea, lográndose en el caso de SPI los mejores valores de "accuracy" (0.565) con $L = 8000$. Nuevamente se debe remarcar en este caso, que un incremento de los valores de L puede generar mejores resultados (0.641 con SPI y $L = 40000$) siendo válidas las mismas consideraciones respecto al costo de clasificación y sobreajuste que se realizaron para el caso del género.

Tabla 1. *Accuracy* para determinación de género en español. *Baseline* = 0.5.

L	KRD	SPI	OOP	RPL	SPI^{re}	OOP^{re}
200	0.544	0.532	0.57	0.473	0.502	0.475
500	0.564	0.546	0.559	0.471	0.501	0.476
1000	0.58	0.553	0.58	0.471	0.51	0.472
2000	0.589	0.58	0.602	0.469	0.531	0.475
3000	0.562	0.571	0.56	0.472	0.526	0.483
4000	0.572	0.572	0.58	0.472	0.536	0.485
6000	0.599	0.581	0.593	0.472	0.538	0.488
8000	0.572	0.57	0.568	0.472	0.54	0.487

Tabla 2. *Accuracy* para determinación de edad en español. *Baseline* = 0.33.

L	KRD	SPI	OOP	RPL	SPI^{re}	OOP^{re}
200	0.435	0.428	0.432	0.313	0.267	0.33
500	0.436	0.445	0.44	0.313	0.264	0.37
1000	0.464	0.488	0.464	0.313	0,267	0.345
2000	0.493	0.497	0.496	0.313	0.267	0.318
3000	0.476	0.526	0.492	0.313	0.267	0.317
4000	0.483	0.531	0.483	0.313	0.268	0.314
6000	0.489	0.544	0.491	0.313	0.269	0.310
8000	0.504	0.565	0.493	0.313	0.27	0.311

5. Conclusiones y Futuras Extensiones

En este trabajo se analizó la viabilidad del uso de enfoques basados en perfiles para la determinación del género y la edad en blogs en español. De acuerdo a nuestro conocimiento, esta es la primera vez que se realiza un estudio de esta naturaleza. De los estudios preliminares, se pudo observar que para el español, el uso de palabras completas es más efectivo que los n-gramas de caracteres y que la determinación conjunta y simultánea de la edad y el género es más efectiva que considerarlas como tareas separadas. Observaciones similares se han realizado para el lenguaje holandés utilizando métodos de clasificación basados en instancias como SVM [8].

Otra observación interesante es que a pesar de su atractivo y efectividad en tareas de AA, el recentrado de perfiles no parece obtener resultados competitivos en esta tarea. Como trabajo a futuro, se planea un análisis más detallado para determinar las causas de este bajo desempeño. Sin embargo, métodos basados en perfiles clásicos como KRD, SPI y OOP obtienen resultados competitivos, comparables a los de otros enfoques más complejos y costosos. En este sentido, SPI a pesar de su simplicidad, ha mostrado resultados prometedores, en particular cuando se incrementa el valor de L para el perfil. Encontrar un valor de L que combine de manera adecuada la precisión en la predicción, bajo costo de clasificación y evite el sobreajuste es un tema de estudio futuro a desarrollarse

cuando se disponga de un conjunto más grande de datos de prueba, como los que serán liberados próximamente como parte del "test set" de la competencia *PAN-PC-2013*.

Es importante observar que más allá de haberse mostrado la viabilidad del uso de este tipo de enfoques en esta clase de problemas, existe un amplio campo para mejorar estos enfoques como pueden ser una selección más cuidadosa de los documentos usados para generar los perfiles de las categorías, el uso de características más informativas para representar los documentos y la combinación de este tipo de métodos con métodos basados en instancia.

Referencias

1. 9th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2013). `http://pan.webis.de/`, 2013.
2. Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52:119–123, 2009.
3. William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.
4. Hugo Jair Escalante, Manuel Montes y Gómez, and Thamar Solorio. A weighted profile intersection measure for profile-based authorship attribution. In *Proceedings of MICAI 2011*, volume 7094, pages 232–243, 2011.
5. Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Source code author identification based on n-gram author profiles. In *Artificial Intelligence Applications and Innovations*, volume 204 of *IFIP*, pages 508–515. Springer US, 2006.
6. Robert Layton, Paul Watters, and Richard Dazeley. Recentred local profiles for authorship attribution. *Natural Language Engineering*, 18:293–312, 2012.
7. India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra Mcbride, and Emma Jakubowski. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122, April 2011.
8. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 37–44, New York, NY, USA, 2011. ACM.
9. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, 2006.
10. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*, 60(3):538–556, 2009.

Una Extensión del FHQT Temporal para Distancias Continuas

Andrés Pascal[1], Anabella De Battista[1], Norma Edith Herrera[2], Gilberto Gutierrez[3]

[1] Dpto. de Sistemas de Información, Universidad Tecnológica Nacional, Entre Ríos, Argentina,
{pascala, debattistaa }@frcu.utn.edu.ar
[2] Dpto. de Informática, Universidad Nacional de San Luis, Argentina,
nherrera@unsl.edu.ar
[3] Universidad del Bio Bio, Facultad de Ciencias Empresariales, Chillán, Chile,
ggutierr@ubiobio.cl

Resumen El modelo de bases de datos métrico-temporal permite abordar aquellas situaciones en las que resulta necesario realizar búsquedas por similitud teniendo en cuenta también la componente temporal. En este artículo presentamos una mejora al índice métrico-temporal *FHQT-Temporal*, que soporta valores continuos de la función de distancia, manteniendo la eficiencia ante cambios de valores del radio de búsqueda e incrementos de los intervalos de tiempo. Además se muestran resultados de la verificación experimental de esta estructura para un conjunto de datos determinado.
Palabras Claves: Espacios Métricos, Bases de Datos Temporales, Índices, Bases de Datos Métrico-Temporales.

1. Introducción

Las bases de datos clásicas se organizan bajo el concepto de búsqueda exacta sobre datos estructurados. Esto significa que la información se organiza en registros los cuales se dividen en campos que contienen valores completamente comparables. Una consulta retorna todos aquellos registros cuyos campos coinciden con los aportados en la consulta (búsqueda exacta). Una característica importante de las bases de datos clásicas es que capturan sólo un estado de la realidad modelizada, usualmente el más reciente. Por medio de las transacciones, la base de datos evoluciona de un estado al siguiente descartando el estado previo.

Actualmente las bases de datos han incluido la capacidad de almacenar datos no estructurados tales como imágenes, sonido, texto, video, datos geométricos, etc. La problemática de almacenamiento y búsqueda en estos tipos de base de datos difiere de las bases de datos clásicas en varios aspectos: los datos no son estructurados por lo que no es posible organizarlos en registros y campos; la búsqueda exacta carece de interés; resulta de interés mantener todos los estados de la base de datos y no sólo el más reciente para poder consultar el intervalo de tiempo de vigencia de los objetos. Es en este contexto donde surgen nuevos modelos de bases de datos.

El modelo de ***espacios métricos*** [3], permite trabajar con objetos no estructurados y realizar búsquedas por similitud sobre los mismos. Un espacio métrico es un par (U, d) donde U es un universo de objetos y $d : U \times U \rightarrow R^+$ es una función de distancia

definida entre los elementos de U que mide la similitud entre ellos. Una de las consultas típicas en este modelo es la búsqueda por rango, denotado por $(q, r)_d$, que consiste en recuperar los objetos de la base de datos que se encuentren como máximo a distancia r de un elemento q dado.

El modelo de ***bases de datos temporales*** [7] incorpora al tiempo como una dimensión, por lo que permite asociar tiempos a los datos almacenados y consultar por los objetos vigentes en un intervalo o en un instante de tiempo dado.

Existen aplicaciones donde resulta de interés realizar búsquedas por similitud teniendo en cuenta también la componente temporal. Es en este ámbito donde surge el ***modelo métrico-temporal***. En este modelo se puede trabajar con objetos no estructurados con tiempos de vigencia asociados y realizar consultas por similitud y por tiempo en forma simultánea.

Varios índices han sido diseñados para resolver consultas métrico-temporales [1,2,5,6]. Todos ellos están basados en el *Fixed Height Queries Tree*, un índice para espacios métricos, y asumen que la función de distancia d devuelve valores discretos. En este trabajo presentamos una extensión del *Fixed Height Queries Tree Temporal(FHQT Temporal)* para distancias continuas.

Este artículo está organizado de la siguiente manera. En la Sección 2 se expone el trabajo relacionado definiendo los conceptos necesarios para la comprensión de este artículo. En la Sección 3 presentamos nuestro aporte definiendo la extensión del *FHQT-Temporal* para distancias continuas. En la Sección 4 presentamos la evaluación experimental y finalizamos en la Sección 5 dando las conclusiones y trabajo futuro.

2. Trabajo Relacionado

2.1. El Modelo Métrico-Temporal

El modelo *métrico-temporal* está orientado a satisfacer búsquedas sobre objetos no estructurados que poseen uno (una sola dimesión temporal) o dos (bitemporal) instantes o intervalos de tiempo asociados y que además no pueden ser recuperados a través de un atributo clave por medio de una búsqueda exacta. Sea U un universo de objetos válidos, se define un Espacio Métrico-Temporal mediante el par (X, d), donde $X = U \times N \times N \times N \times N$ y d es la función de distancia $d : U \times U \rightarrow R^+$. Cada elemento $x \in X$ es una 5-upla $(o, t_{vi}, t_{vf}, t_{ti}, t_{tf})$, donde o es un objeto (una huella digital, una imagen, un sonido, etc), $[t_{vi}, t_{vf}]$ es el intervalo de validez de o en la realidad y $[t_{ti}, t_{tf}]$ el intervalo de tiempo transaccional asociado. Por simplicidad, se definen todos los tiempos como valores pertenecientes al conjunto N. Estos valores pueden ser fechas, horas, etc., pero en cualquier caso se pueden representar mediante números naturales. La función de distancia d mide la disimilitud entre dos objetos y cumple con las propiedades de toda métrica, es decir, positividad, simetría, reflexividad y desigualdad triangular [3].

Formalmente una *consulta métrico-temporal* se define como una 4-upla $(q, r, t_{iq}, t_{fq})_d$, tal que:

$$(q, r, t_{iq}, t_{fq})_d = \{o/(o, t_{io}, t_{fo}) \in X \wedge d(q, o) \leq r \wedge (t_{io} \leq t_{fq}) \wedge (t_{iq} \leq t_{fo})\}$$

Una forma trivial de resolver una consulta métrico-temporal, sin realizar una búsqueda exhaustiva en la bases de datos, es construir un índice métrico agregándole a cada

objeto un intervalo temporal que represente la vigencia del mismo. Luego, ante una consulta $(q, r, t_{iq}, t_{fq})_d$ en primer lugar se utilizará el índice métrico para descartar los objetos obj que están a distancia mayor que r de q; y posteriormente se realizará un recorrido del conjunto de elementos no descartados en el primer paso para determinar qué objetos conforman la respuesta a la consulta, que serán aquellos cuyo intervalo de vigencia que se superpone con $[t_{iq}, t_{fq}]$.

Varios índices métrico-temporales se han propuesto en este ámbito: el *Pivot-FHQT* [1], el *Historical-FHQT* [2], el *Event-FHQT* [5] y el *FHQT-Temporal* [6]; todos ellos han tomado como base el Fixed Height Queries Tree[3], un índice para espacios métricos. Todos estos índices asumen que la función de distancia retorna valores discretos.

2.2. FHQT-Temporal

El *FHQT-Temporal* es un FHQT al cual se le agrega un intervalo de tiempo en cada nodo del árbol. Este intervalo representa el período de tiempo de vigencia de todos los objetos del subárbol cuya raíz es dicho nodo. En cada nodo hoja, este intervalo es el período total de vigencia de los objetos que contiene. Para un nodo interior, el intervalo se calcula tomando el tiempo inicial mínimo y el tiempo final máximo de sus hijos.

Este índice permite resolver consultas por similitud puras, temporales puras y métrico-temporales. Como se basa en un FHQT, solo permite funciones de distancia discretas.

Formalmente, un *FHQT-Temporal* es un árbol en el cual un nodo interior V es una 3-upla (t_{ini}, t_{fin}, $\{(d_1, h_1), (d_2, h_2),..., (d_m, h_m)\}$) donde:

- $h_1..h_m$ son los m hijos del nodo V,
- las d_i , para $i = 1..m$, son las distancias entre el pivote correspondiente al nivel de V y los objetos contenidos en las hojas de los subárboles de h_i.
- los dos primeros componentes de la 3-upla, t_{ini} y t_{fin}, se definen de la siguiente manera: t_{ini} = $min_{j=1..m}(t_{ini}(h_j))$, y t_{fin} = $max_{j=1..m}$ $(t_{fin}(h_j))$.

Las hojas del *FHQT-Temporal* tienen una estructura similar; están representadas por una 3-upla (t_{ini}, t_{fin}, $\{e_1, e_2,..., e_l\}$), donde:

- los e_i para $i = 1..l$ son los l elementos que contiene la hoja, que a su vez están formados por tres componentes: el objeto o, el tiempo inicial del mismo t_{io}, y su tiempo final t_{fo}.
- los valores t_{ini}, t_{fin} poseen el mismo significado que para los nodos interiores, pero aplicados a los elementos e_i.

En la Figura 1 se muestra el esquema genérico del *FHQT-Temporal*. La estructura es dinámica, permitiendo tanto altas como bajas ya sea de instantes o intervalos contenidos en el intervalo que el índice posee hasta el momento, como de objetos con tiempos fuera de éste.

Cuando se realiza una consulta métrico-temporal, se procede de la siguiente manera: en cada nivel del árbol se seleccionan los subárboles hijos del nodo que se está procesando, cuyos intervalos temporales se intersectan con el intervalo o instante de la consulta. De éstos, posteriormente se eligen los que cumplen con la restricción de similitud tomando en cuenta la firma de la consulta y el radio de búsqueda. Este procedimiento

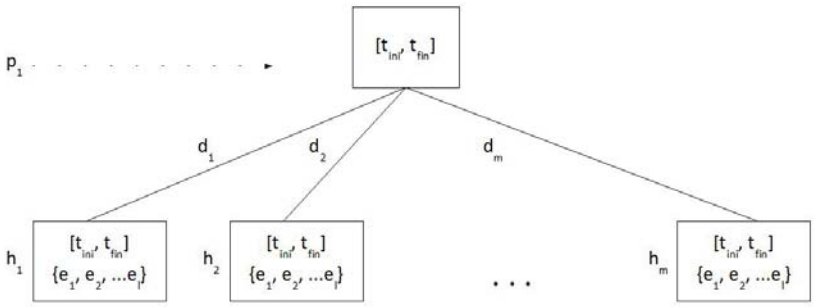


Figura 1. *Esquema del FHQT-Temporal*

se repite hasta llegar al último nivel. Para cada hoja no descartada, luego de verificar la superposición temporal, se realiza una búsqueda secuencial sobre todos los elementos contenidos en las mismas, comparando tanto el aspecto temporal como la distancia de cada elemento a la consulta.

3. Extensión del FHQT-Temporal para Distancias Continuas

El *FHQT*$^+$*-Temporal* es una variante del *FHQT-Temporal* generalizada que soporta valores continuos de la función de distancia. Para ello, en lugar de asociar un número natural a cada hijo de un nodo, se asocian dos intervalos de valores de distancias. El primer intervalo representa el rango máximo correspondiente a la rama, mientras que el segundo (incluido en el anterior) constituye el rango actual de valores, es decir, el intervalo formado por el mínino y el máximo valor de distancia del pivote a los objetos contenidos en las hojas del subárbol. Los intervalos máximos son constantes y se calculan cuando se construye el árbol en base al histograma de distancias, de tal manera de que el árbol tenga una alta probabilidad de quedar balanceado. Los intervalos actuales son variables y se van actualizando de acuerdo a los objetos que se insertan en la estructura. Para determinar en qué rama se agrega un nuevo elemento, se utilizan los intervalos máximos y ante una consulta sólo se usan los actuales. Al utilizar estos últimos intervalos en las búsquedas, se incrementa la capacidad de filtrado por similitud, ya que los intervalos son más pequeños y quedan espacios vacíos entre intervalos consecutivos, como veremos más adelante.

Un *FHQT*$^+$*-Temporal* es un árbol r-ario donde el valor de r es un parámetro que se define en forma previa a su construcción, normalmente en base a la distribución del histograma de distancias. Formalmente, es un árbol donde cada nodo interior V es una 3-upla (t_{ini}, t_{fin}, $\{(int_{x1}, int_{a1}, h_1), (int_{x2}, int_{a2}, h_2),..., (int_{xm}, int_{am}, h_m)\}$) donde:

- $h_1..h_m$ son los m hijos del nodo V,
- los int_{xi} , para $i = 1..m$, son los intervalos *máximos* de distancias entre el pivote correspondiente al nivel de V y los objetos que pueden pertenecer a las hojas de los subárboles de h_i.

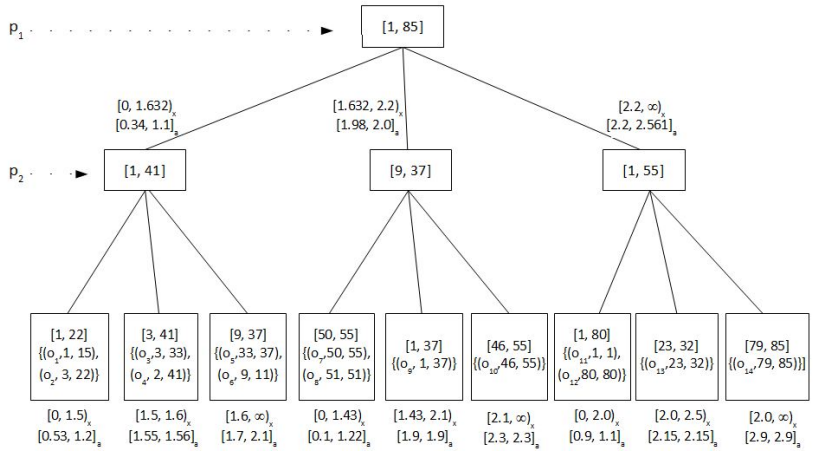


Figura 2. *Ejemplo de un FHQT$^+$-Temporal*

- los int_{ai} , para $i = 1..m$, son los intervalos *actuales* de distancias entre el pivote correspondiente al nivel de V y los objetos contenidos actualmente en las hojas de los subárboles de h_i.
- los dos primeros componentes de la 3-upla, t_{ini} y t_{fin}, se definen de la siguiente manera: t_{ini} = $min_{j=1..m}(t_{ini}(h_j))$, y t_{fin} = $max_{j=1..m}\ (t_{fin}(h_j))$.

Las hojas del *FHQT$^+$-Temporal* se definen de la misma forma que para el *FHQT-Temporal*

Para calcular los intervalos máximos correspondientes al nodo raíz del árbol, se toma una muestra de la base de datos, se calcula el histograma de distancias y se divide el espacio en r intervalos, de tal manera de que cada uno de ellos posea la misma cantidad (± 1) de elementos. Luego para cada nodo hijo se procede de la misma manera, pero considerando solo los elementos de la muestra que fueron asignados a dicho nodo. De esta manera todos los nodos interiores tendrán exactamente r hijos.

Una vez definidos los rangos, se realiza la inserción de los elementos, actualizando los intervalos actuales. Sea o el objeto a insertar, v el nodo donde se quiere insertar el objeto, $[d_{xi}, d_{xf})$ y $[d_{ai}, d_{af}]$ los intervalos máximo y actual asociados al nodo, y p el pivote del nivel, primero se verifica que $d(p, o) \in [d_{xi}, d_{xf})$ y si esto se cumple, se actualiza el intervalo actual haciendo $d_{ai} := min(d_{ai}, d(p, o))$ y $d_{af} := max(d_{af}, d(p, o))$. El aspecto temporal se procesa de la misma manera que en el *FHQT-Temporal*.

Ante una consulta, para visitar un nodo se comprueba que la distancia de la consulta al pivote del nivel pertenezca al intervalo actual asociado al nodo mas menos el radio de búsqueda, es decir, que $f_n \in [d_{ai} - r, d_{af} + r]$. En la Figura 2 se muestra un ejemplo del *FHQT$^+$-Temporal*. Es interesante notar que los intervalos actuales correspondientes a los nodos de un mismo nivel, usualmente no cubren todo el espacio posible. Al reducir el tamaño de estos rangos, aumenta la probabilidad de que una rama se descarte ya que la comprobación anterior se realiza sobre un intervalo más pequeño. Por ejemplo, si el radio de búsqueda es $0{,}3$ y la distancia entre la consulta y el pivote del primer nivel es

1,5, si se utilizan los rangos máximos se deben procesar el primer y segundo hijo del nodo raíz, ya que $1{,}5 \in [0-0{,}3\ ,\ 1{,}632+0{,}3)$ y $1{,}5 \in [1{,}632-0{,}3\ ,\ 2{,}2+0{,}3)$, mientras que usando los rangos actuales ambos se descartan porque $1{,}5 \notin [0-0{,}34\ ,\ 1{,}1+0{,}3]$ y $1{,}5 \notin [1{,}98-0{,}3\ ,\ 2{,}0+0{,}3]$.

Este índice permite resolver consultas por similitud puras, temporales puras y métrico-temporales con funciones de distancia tanto continuas como discretas.

4. Evaluación Experimental

Debido a que el modelo métrico-temporal es relativamente reciente, no existen aún bases de datos disponibles para realizar experimentos, por lo que se optó por adaptar la base $NASA$ [4] que es frecuentemente utilizada por investigadores del área de espacios métricos (disponible para su descarga en *http://www.sisap.org/library/dbs/vectors/*), para la determinación experimental de la eficiencia de esta nueva estructura ante consultas métrico-temporales.

La base de datos $NASA$ es un conjunto de 40.150 vectores 20-dimensionales de números reales, que representan características de imágenes obtenidas por la $NASA$.

Partiendo de este conjunto de datos se generó la base de datos métrico-temporal *NASA*^{MT} asignando a cada vector un identificador y un intervalo de vigencia, que indica el período de validez del objeto. El intervalo total considerado fue [1, 1000]. Luego, mediante un proceso aleatorio se generaron lotes de 1.000, 5.000, 10.000, 20.000 y 30.000 elementos.

Una vez construido el índice, se seleccionaron al azar 100 objetos de la base de datos *NASA*^{MT} y se generaron cuatro lotes de consultas métrico-temporales mediante la asignación en forma aleatoria de intervalos/instantes de tiempo. Uno de los lotes para cada base de datos se compuso solamente de consultas instantáneas y los demás fueron construídos asociándoles intervalos correspondientes al 10 %, 25 % y 50 % del intervalo total ($[1, 1000]$).

Para completar los parámetros requeridos en las consultas, se definieron tres radios de búsquedas distintos para cada base de datos, que devuelven en promedio aproximadamente el 1 %, 5 % y 10 % de los objetos contenidos ante las consultas por similitud de los lotes definidos anteriormente. Estos radios fueron calculados experimentalmente y son los siguientes: 0,453; 0,69855 y 0,8275.

Como función de distancia se utilizó la distancia euclidiana que es la medida utilizada usualmente sobre esta bases de datos para realizar pruebas por similitud. En estas pruebas sólo se tomó en cuenta como variable de costo la cantidad de evaluaciones de la función de distancia ya que la estructura se mantuvo en memoria principal.

4.1. FHQT$^+$-Temporal: Análisis de los Resultados Obtenidos

En esta sección se presentan, grafican y analizan los resultados obtenidos para el *FHQT*$^+$*-Temporal* en comparación con la solución trivial que utiliza un FHQT como índice métrico donde cada objeto tiene su intervalo de vigencia asociado. En la solución triial, primero se busca por similitud y para cada objeto resultante de esta búsqueda, se usa su intevalo de vigencia asociado para determinar si forma o no parte de la respuesta.

Cabe notar que cualquiera de dichas soluciones tiene, como mínimo, el costo correspondiente a un índice métrico.

Variación del Costo en Función del Tamaño de la Base de Datos En los gráficos de la Figura 3 se muestran las curvas de costos correspondientes a los lotes de consultas instantáneas y 50 % del intervalo respectivamente, en comparación con la solución trivial. El eje x indica la cantidad de elementos de la base de datos NASA^{MT} y el eje y, el promedio de evaluaciones de la función de distancia.

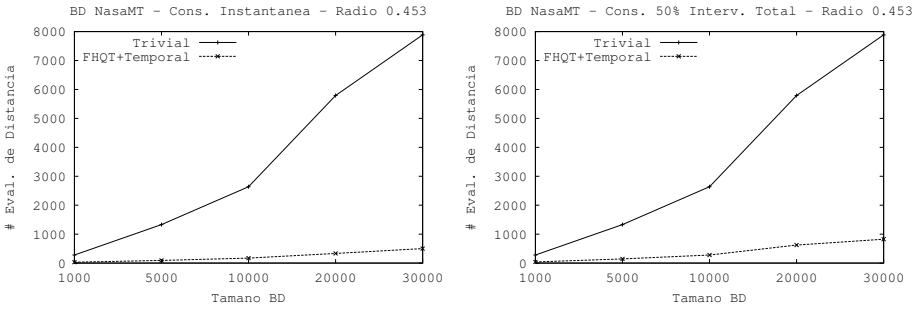


Figura 3. *Costo consultas métrico-temporales mediante un FHQT$^+$-Temporal a la base de datos NASA^{MT}, en función del tamaño)*

Las curvas muestran en ambos casos que ante el aumento de la cantidad de elementos del conjunto de datos consultado, el costo se incrementa de forma similar a una curva lineal aunque los factores por los cuales se multiplica son muy distintos. Claramente se ve que el *FHQT$^+$-Temporal* supera ampliamente la performance de la solución trivial. En el mejor de los casos –cuando el tamaño es el mayor y las consultas son instantáneas–, su costo es de sólo el 7,89 % del correspondiente a la solución trivial, y en el peor de los casos, un 12,6 %. Es importante notar que el costo de la solución trivial no varía en función del tamaño del intervalo de tiempo, por lo cual en ambos gráficos la curva es la misma.

El *porcentaje de evaluaciones*, medido como la cantidad de objetos resultantes sobre el costo promedio de las consultas, fue del $0{,}1$ % al $1{,}0$ % para la solución trivial, es decir que por cada elemento resultante tuvieron que ser evaluados 1000 objetos en el peor de los casos y 100 en el mejor caso. Para el *FHQT$^+$-Temporal* este porcentaje se eleva a $0{,}7$ % y $9{,}4$ % respectivamente, correspondientes a 143 y 11 evaluaciones de la función de distancia por cada objeto resultante. En ambos casos, las mejoras no son producidas por la variación de la cantidad de elementos, sino por la modificación del radio e intervalo de búsqueda.

El *FHQT$^+$-Temporal* supera claramente en eficiencia a la solución trivial en todos los casos, por lo cual en los apartados siguientes sólo se analizan las variaciones de este índice respecto a los distintos parámetros.

Variación del Costo en Función del Radio de Búsqueda En la Figura 4 se presentan dos gráficas de costo del *FHQT*$^+$*-Temporal* en función del radio de búsqueda. La primera corresponde a consultas instantáneas y la segunda a intervalos de tiempo con tamaño promedio igual al 50 % del total.

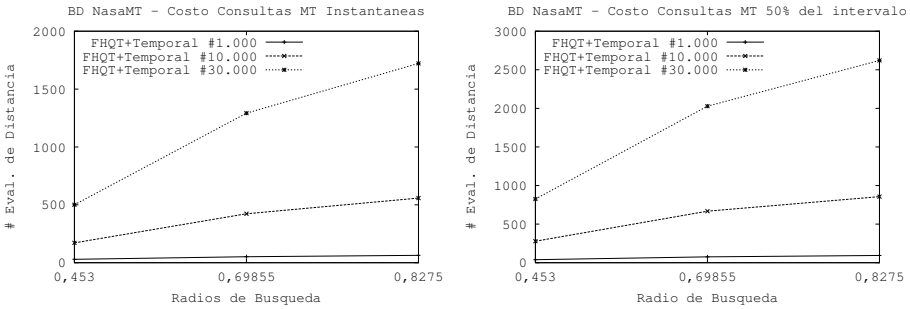


Figura 4. *Costo consultas métrico-temporales mediante un FHQT*$^+$*-Temporal a la base de datos NASA*^{MT}*, en función del radio*

Como es lógico, la cantidad de evaluaciones de la función de distancia se incrementa considerablemente al aumentar el radio de búsqueda. Sin embargo, el porcentaje de evaluaciones mejora sustancialmente. Este porcentaje varía entre $0{,}7$ y $0{,}9$ para el radio menor, y es alrededor de 9 veces más grande para el mayor radio. Esto significa que la eficiencia del *FHQT*$^+$*-Temporal* aumenta cuando se incrementa el radio de búsqueda. En todo índice métrico, es natural que el porcentaje de evaluaciones en algún momento aumente al consultar con mayores radios debido a que la cantidad de resultados también es mayor. En el caso extremo, cuando se consulta con un radio que incluye a todos los elementos de la base de datos, este porcentaje es cercano a 100. Sin embargo, en una base métrico-temporal, si sólo se aumenta el radio y el intervalo de tiempo de la consulta permanece constante, puede ser que la cantidad de resultados sea la misma ya que no todos los elementos cumplirán la restricción temporal.

En el *FHQT*$^+$*-Temporal*, cuando se aumenta el radio de búsqueda las restricciones temporales se hacen más importantes para el proceso de descarte de elementos, es decir que la cantidad de elementos que cumplen con la condición de similitud es mayor, pero la cantidad de elementos que cumplen con la restricción temporal se mantiene igual. Esta es una de las causas del aumento del porcentaje de evaluaciones.

Variación del Costo en Función de la Amplitud del Intervalo de Tiempo Para evaluar la influencia de las variaciones de la amplitud del intervalo temporal sobre el costo de las consultas se presentan los gráficos de la Figura 5. En el eje X se ubican en primer lugar las consultas instantáneas y a continuación los intervalos temporales correspondientes al $10\,\%$, $25\,\%$ y $50\,\%$ del intervalo total. Los valores del eje Y representan las cantidades promedio de evaluaciones de la función de distancia para los lotes de 100

consultas. El radio de consulta correspondiente al primer gráfico es $0{,}453$ y el del segundo $0{,}8275$.

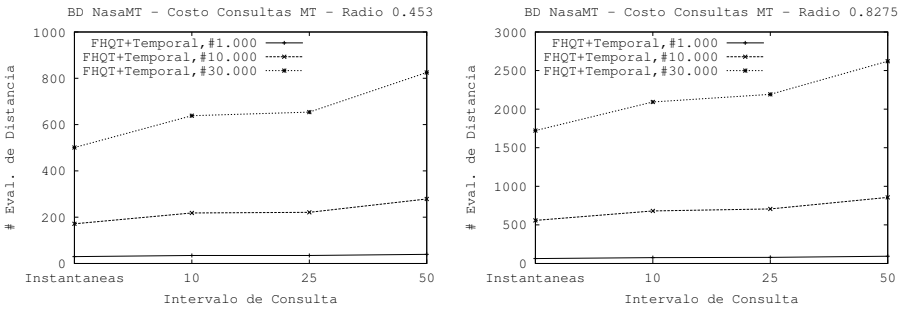


Figura 5. *Costo consultas métrico-temporales mediante un FHQT$^+$-Temporal a la base de datos NASA^{MT}, en función del intervalo temporal*

Como se ve en ambos gráficos, el costo de las consultas que toman el 50 % del intervalo total es entre un 50 % y un 70 % mayor que el de las consultas instantáneas. Por otro lado, la cantidad de elementos resultantes es alrededor de dos veces mayor (lo que es coherente con el aumento del intervalo temporal, ya que los objetos se encuentran uniformemente distribuidos en cuanto al tiempo). Por esta razón el porcentaje de evaluaciones aumenta también, entre un 30 % y un 40 %.

5. Conclusiones y Trabajo Futuro

En este trabajo presentamos una extensión de un método de acceso métrico-temporal que soporta valores continuos de la función de distancia, el *Fixed Height Queries Tree$^+$-Temporal(FHQT$^+$-Temporal)*, que permite resolver eficientemente consultas métrico-temporales. Este índice permite resolver consultas por similitud puras, temporales puras y métrico-temporales con funciones de distancia tanto continuas como discretas. Los experimentos han mostrado que el FHQT$^+$-Temporal tiene en la totalidad de los casos un costo menor que la solución trivial, y son notables las mejoras respecto al porcentaje de evaluaciones del índice cuando se incrementan los radios y los intervalos de consulta.

Actualmente estamos trabajando en la extensión de otros índices métrico-temporales para distancias continuas.

Referencias

1. A. De Battista, A. Pascal, N. Herrera, and G. Gutierrez. Metric-temporal access methods. *Journal of Computer Science & Technology,*, 10(2):54–60, 2010.

2. De Battista, A. Pascal, G. Gutierrez, and N. Herrera. Un nuevo índice métrico-temporal: el historical fhqt. In *Actas del XIII Congreso Argentino de Ciencias de la Computación*, Corrientes, Agentina, 2007.
3. Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM Comput. Surv.*, 33(3):273–321, 2001.
4. K. Figueroa, G. Navarro, and E. Chávez. Metric spaces library, 2007. Available at `http://www.sisap.org/Metric_Space_Library.html`.
5. A. Pascal, A. De Battista, G. Gutierrez, and N. Herrera. Indice metrico-temporal event-fhqt. In *Actas del XIIII Congreso Argentino de Ciencias de la Computación*, La Rioja, Argentina, 2008.
6. A. Pascal, De Battista, G. Gutierrez, and N. Herrera. Procesamiento de consultas métrico-temporales. In *XXIII Conferencia Latinoamericana de Informática*, pages 133–144, Costa Rica, 2007.
7. B. Salzberg and V. J. Tsotras. A comparison of access methods for temporal data. *ACM Computing Surveys*, 31(2), 1999.

New Deletion Method for Dynamic Spatial Approximation Trees

Fernando Kasián, Verónica Ludueña, Nora Reyes, and Patricia Roggero

Departamento de Informática, Universidad Nacional de San Luis,
San Luis, Argentina
{fkasian,vlud,nreyes,proggero}@unsl.edu.ar

Abstract. The Dynamic Spatial Approximation Tree (*DSAT*) is a data structure specially designed for searching in metric spaces. It has been shown that it compares favorably against alternative data structures in spaces of high dimension or queries with low selectivity. The *DSAT* supports insertion and deletions of elements. However, it has been noted that eliminations degrade the structure over time. In [8] is proposed a method to handle deletions over the *DSAT*, which shown to be superior to the former in the sense that it permits controlling the expected deletion cost as a proportion of the insertion cost.
In this paper we propose and study a new deletion method, based on the deletions strategies presented in [8], which has demonstrated to be better. The outcome is a fully dynamic data structure that can be managed through insertions and deletions over arbitrarily long periods of time without any reorganization.

Keywords: multimedia databases, metric spaces, similarity search

1 Introduction

"Proximity" or "similarity" searching is the problem of looking for objects in a set close enough to a query under a certain (expensive to compute) distance. Similarity search has become a very important operation in applications that deal with unstructured data sources. For example, multimedia databases manage objects without any kind of structure, such as images, fingerprints or audio clips. This has applications in a vast number of fields. Some examples are non–traditional databases, text searching, information retrieval, machine learning and classification, image quantization and compression, computational biology, and function prediction. All those applications can be formalized with the *metric space model* [3]. That is, there is an universe $\mathcal{U}$ of objects, and a positive real valued distance function $d : \mathcal{U} \times \mathcal{U} \longrightarrow \mathbb{R}^+$ defined among them. This distance may (and ideally does) satisfy the three axioms that make the set a metric space: *strict positiveness*, *symmetry*, and *triangle inequality*. The smaller the distance between two objects, the more "similar" they are. We have a finite database $S \subseteq \mathcal{U}$, which is a subset of the universe and can be preprocessed. Later, given a new object from the universe (a *query* q), we must retrieve all similar elements found in the database. There are two typical queries of this kind:

Range query: retrieve all elements within distance r to q in S.
Nearest neighbor query (k-NN): retrieve the k closest elements to q in S.

The distance is considered expensive to compute. Hence, it is customary to define the complexity of search as the number of distance evaluations performed. We consider the number of distance evaluations instead of the CPU time because the CPU overhead over the number of distance evaluations is negligible in the *DSAT*. In this paper we are devoted to range queries. In [5] is shown how to build an nearest neighbors algorithm range-optimal using a range algorithm, so we can restrict our attention to range queries.

Proximity search algorithms build an *index* of the database and perform queries using this index, avoiding the exhaustive search. For general metric spaces, there exist a number of methods to preprocess the database in order to reduce the number of distance evaluations [3]. All those structures work on the basis of discarding elements using the triangle inequality, and most use the classical divide-and-conquer approach. (which is not specific of metric space searching).

The Spatial Approximation Tree (*SAT*) is a proposed data structure of this kind [6, 7], based on a concept: approach the query spatially. It has been shown that the *SAT* gives better space-time tradeoffs than the other existing structures on metric spaces of high dimension or queries with low selectivity [7], which is the case in many applications. The *SAT*, however, has some important weaknesses: it is relatively costly to build in low dimensions; in low dimensions or for queries with high selectivity (small r or k), its search performance is poor when compared to simpler alternatives; and it is a static data structure: once built, it is hard to add/delete elements to/from it. These weaknesses make the *SAT* unsuitable for important applications such as multimedia databases.

The *DSAT* is a dynamic version of the *SAT* and overcomes its drawbacks. The dynamic *SAT* can be built incrementally (i.e., by successive insertions) at the same cost of its static version, and the search performance is unaffected. At first, the *DSAT* supports insertion and deletions of elements. However, that deletions degrade the structure over time, so in [8] was presented a deletion algorithm that does not degrade the search performance over time. This algorithm yielded better tradeoffs between search performance and deletion cost. In this paper we present another alternative method to delete an element of the *DSAT* which obtains better costs than methods described in [8].

Full dynamism is not so common in metric data structures [3]. While permitting efficient insertions is quite usual, deletions are rarely handled. In several indexes one can delete some elements, but there are selected elements that cannot be deleted at all. This is particularly problematic in the metric space scenario, where objects could be very large (e.g., images) and deleting them physically may be mandatory. Our algorithms permit deleting any element from a *DSAT*.

This paper is organized as follows: In Section 2 we give a description of the *DSAT*. Section 3 presents our new improved deletion method, and Section 4 contains the empirical evaluation of our proposal. Finally, in Section 5 we conclude and discuss about possible extensions for our work.

2 Dynamic Spatial Approximation Trees

In this section we briefly describe dynamic *SAT* (*DSAT* for short), in particular the version called *timestamp with bounded arity* presented in [8] as the better option to build incrementally the index, without any reconstruction after each insertion. To construct *DSAT* [8] incrementally a maximum tree arity is fixed, and also a timestamp of the

Algorithm 1 Insertion of a new element x into a *DSAT* with root a.

```
Insert(Node a, Element x)
1. R(a) ← max(R(a), d(a,x))
2. c ← argmin_{b∈N(a)} d(b,x)
3. If d(a,x) < d(c,x) ∧ |N(a)| < MaxArity Then
4.    N(a) ← N(a) ∪ {x}
5.    N(x) ← ∅,  R(x) ← 0
6.    time(x) ← CurrentTime
7.    CurrentTime ← CurrentTime + 1
8. Else Insert(c,x)
```

insertion time of each element is kept. Each node a in the tree is connected to its children, which form a set of elements called $N(a)$, the *neighbors* of a. When inserting a new element x, its point of insertion is found by beginning from the tree root a and pe-rforming the following procedure. The element x is added to $N(a)$ (as a new leaf node) if (1) x is closer to a than to any element $b \in N(a)$, and (2) the arity of node a, $|N(a)|$, is not already maximal. Otherwise x is forced to choose the closest neighbor in $N(a)$ and keep walking down the tree in a recursive manner, until we reach a node a such that x is closer to a than any $b \in N(a)$ and the arity of node a is not maximal (this eventually occurs at a tree leaf). At this point x is added at the end of the list $N(a)$, the current timestamp is put to x and the current timestamp is incremented. The following information is kept in each node a of the tree: the set of neighbors $N(a)$, the timestamp $time(a)$ of the insertion time of the node, and the covering radius $R(a)$ with the distance between a and the farthest element in the subtree of a.

Note that by reading neighbors from left to right timestamps increase. It also holds that the parent is always older than its children. The *DSAT* can be built by starting with a first single node a where $N(a) = \emptyset$ and $R(a) = 0$, and then performing successive insertions. Algorithm 1 gives the insertion process.

2.1 Searching

The idea for range searching is to replicate the insertion process of relevant elements. That is, the process act as if it wanted to insert q but keep in mind that relevant elements may be at distance up to r from q, so in each decision for simulating the insertion of q a tolerance of $\pm r$ is permitted, so that it may be that relevant elements were inserted in different children of the current node, and backtracking is necessary.

Two facts have to be considered. The first is that, when an element x was inserted, a node a in its path may not have been chosen as its parent because its arity was already maximal. So, at query time, instead of choosing the closest to x among $\{a\} \cup N(a)$, it may have chosen only among $N(a)$. Hence, the minimization is performed only among elements in $N(a)$. The second fact is that, at the time x was inserted, elements with higher timestamp were not yet present in the tree, so x could choose its closest neighbor only among elements older than itself.

A better use of the timestamp information is made in order to reduce the work done inside older neighbors. Say that $d(q, b_i) > d(q, b_{i+j}) + 2r$. The process enters into the subtree of b_i anyway because b_i is older. However, only the elements with

Algorithm 2 Searching for q with radius r in a *DSAT* rooted at a.

```
RangeSearch(Node a, Query q, Radius r, Timestamp t)
 1. If time(a) < t  ∧  d(a,q) ≤ R(a) + r Then
 2.   If d(a,q) ≤ r Then Report a
 3.   d_min ← ∞
 4.   For b_i ∈ N(a) Do   /* in ascending timestamp order */
 5.     If d(b_i,q) ≤ d_min + 2r Then
 6.       t' ← min{t} ∪ {time(b_j), j > i ∧ d(b_i,q) > d(b_j,q) + 2r}
 7.       RangeSearch(b_i, q, r, t')
 8.       d_min ← min{d_min, d(b_i,q)}
```

timestamp smaller than that of b_{i+j} should be considered when searching inside b_i; younger elements have seen b_{i+j} and they cannot be interesting for the search if they are inside b_i. As parent nodes are older than their descendants, as soon as a node inside the subtree of b_i with timestamp larger than that of b_{i+j} is found the search in that branch can stop, because all its subtree is even younger.

Algorithm 2 shows the process to perform range searching. Note that, except in the first invocation, $d(a, q)$ is already known from the invoking process.

2.2 Deletions

To delete an element x, the first step is to find it in the tree. Unlike most classical data structures, doing this is not equivalent to simulating the insertion of x and seeing where it leads us to in the tree. The reason is that the tree was different at the time x was inserted. If x were inserted again, it could choose to enter a different path in the tree, which did not exist at the time of its first insertion.

An elegant solution to this problem is to perform a range search with radius zero, that is, a query of the form $(x, 0)$. This is reasonably cheap and will lead us to all the places in the tree where x could have been inserted.

On the other hand, whether this search is necessary is application dependent. The application could return a handle when an object was inserted into the database, and therefore this search would not be necessary. This handle can contain a pointer to the corresponding tree node. Adding pointers to the parent in the tree would permit to locate the path for free (in terms of distance computations). Hence, in which follows, the location of the object is not considered as part of the deletion problem, although it has shown how to proceed if necessary.

Several alternatives to delete elements from *DSAT* were studied in [8]. From the beginning they discarded the trivial option of marking the element as deleted without actually deleting it. As explained, this is likely to be unacceptable in most applications. It is assumed that the element has to be physically deleted. It may, if desired, keep its node in the tree, but not the object itself. It should be clear that a tree leaf can always be deleted without any complication, so the focus is on how to remove internal tree nodes.

There are several proposed methods to delete an element from a *DSAT*, but in [8] the authors showed that the best option is based in *ghost hyperplanes*. This technique is inspired on an idea presented in [10] for dynamic *gna–trees* [2], called *ghost hyperplanes*. This method replaces the element being deleted by a leaf, which is easy to

Algorithm 3 Deleting x from a *DSAT*, finding a substitute in the leaves of its subtree.

DeleteGH1(Node x)
1. $b \leftarrow parent(x)$
2. If $N(x) \neq \emptyset$ Then
3. $y \leftarrow$ **FindSubstituteLeaf** (x)
4. $d_g(x) \leftarrow d_g(x) + d(x, y)$
5. Copy object of y into node x
6. Else $N(b) \leftarrow N(b) - \{x\}$

FindSubstituteLeaf(Node x): Node
1. $y \leftarrow x$
2. While $N(y) \neq \emptyset$ Do
3. $x \leftarrow y$
4. $y \leftarrow \text{argmin}_{c \in N(b)} d(c, x)$
5. $N(x) \leftarrow N(x) - \{y\}$
6. Return y

delete. This way rebuilding is not necessary, but in exchange some tolerance must be considered when entering the replaced node at search time.

Remind that the neighbors of a node b in the *DSAT* partition the space in a Voronoi-like fashion, with hyperplanes. If element y replaces a neighbor x of b, the hyperplanes will be shifted (slightly, if y is close to x). We can think of a "ghost" hyperplane, corresponding to the deleted element x, and a real one, corresponding to the new element y. The data in the tree is initially organized according to the ghost hyperplane, but incoming insertions will follow the real hyperplane. A search must be able to find all elements, inserted before or after the deletion of x.

For this sake, we have to maintain a tolerance $d_g(x)$ at each node x. This is set to $d_g(x) = 0$ when x is first inserted. When x is deleted and the content of its node is replaced by y, we will set $d_g(x) = d_g(x) + d(x, y)$ (the node is still called x although its object is that of y). Note that successive replacements may shift the hyperplanes in all directions so the new tolerance must be added to previous ones.

At search time, we have to consider that each node x can actually be offset by $d_g(x)$ when determining whether or not we must enter a subtree. Therefore, we wish to keep $d_g()$ values as small as possible, that is, we want to find replacements that are as close as possible to the deleted object. When node x is deleted, we have to look for a substitute in its subtree to ensure that we reduce the problem size.

Choosing a leaf substitute We descend in the subtree of x by the children closest to x all the time. When it reach a leaf y, it disconnect y from the tree and put y into the node of x, retaining the original timestamp of x. Then, the d_g value of the node is updated.

Choosing a neighbor substitute We select y as the closest to x among $N(x)$ and copy object y into the node of x as above. If the former node of y was a leaf it delete it and finish. Otherwise we recursively continue the process at that node. So, we turn to *ghost* all the nodes in the path from x to a leaf of its subtree, following the closest neighbors. In exchange, the $d_g()$ values should be smaller.

Choosing the nearest-element substitute We select y as the closest element to x among all the elements in the subtree of x and copy object y into the node of x as above. If the former node of y was a leaf we delete it and finish. Otherwise, we recursively continue the process at that node. Therefore, we turn to *ghost* some nodes in the path from x to a leaf of its subtree, following the nearest elements. The $d_g()$ values should be smaller than with the other alternatives.

Algoritms 3, 4, and 5 detail these three deletion methods.

Algorithm 4 Deleting x from a *DSAT*, choosing its replacement among its neighbors.

```
DeleteGH2(Node x)
1. b ← parent(x)
2. If N(x) ≠ ∅ Then
3.    y ← argmin_{c∈N(x)} d(c,x), d_g(x) ← d_g(x) + d(x,y)
4.    Copy object of y into node x
5.    DeleteGH2 (y)
6. Else N(b) ← N(b) − {x}
```

Algorithm 5 Deleting x from a *DSAT*, choosing its replacement as its nearest element.

```
DeleteGH3(Node x)
1. b ← parent(x)
2. If N(x) ≠ ∅ Then
3.    y ← NNsearch(x,x,1), d_g(x) ← d_g(x) + d(x,y)
4.    Copy object of y into node x
5.    DeleteGH3 (y)
6. Else N(b) ← N(b) − {x}
```

Thus, for a permanent regime that includes deletions, we must periodically get rid of ghost hyperplanes and reconstruct the tree to delete them. Just as with fake nodes [8], when we rebuild the subtree we get rid of all the ghost hyperplanes that are inside it. We set a maximum allowable proportion α of ghost hyperplanes, and rebuild the tree when this limit is exceeded.

3 A New Deletion Method

In [8] it is concluded that the methods with the best performance during deletions use ghosts hyperplanes. Moreover, these methods have the possibility of using the parameter α to control the deletion average cost. Our new proposal to delete an element x is based on the best strategies presented in [8]. Therefore, this new proposed method is also based on the idea presented in [10], which use *ghost hyperplanes*.

We believe that the way to achieve a good tradeoff between the number of hyperplanes and the displacement of each d_f can be obtained by replacing the deleted element x with the leaf of his subtree whose distance is minimal; i. e. *the closest leaf in the complete subtree of* x. Therefore, with each deletion only one new ghost hyperplane appears and the displacement of this ghost hyperplane, although it is not necessarily the smallest one possible, is expected to be fairly close to it.

It is possible to notice, considering the presented algorithms in Section 2.2, that it is only needed to change the process invoked as **FindSubstituteLeaf**(x). This new algorithm has to choose the closest element to x between all the leaves in the subtree of x. Therefore, **DeleteGH1**(Node x) is similar to **DeleteGH4**(Node x) because only one ghost hyperplane appears after deletion. The Algorithm 6 shows this situation completely. In the function **FindSubstituteNNLeaf** all the leaves of the subtree of x are recovered in the set L of pairs (z,t), where z is a leaf of the subtree of x and t is his father. Q is a queue of elements to be used as an auxiliary data structure in a *traversal*

Algorithm 6 Deleting x from a *DSAT*, finding a substitute as the *closest leaf*.

```
DeleteGH4(Node x)
 1. b ← parent(x)
 2. If N(x) ≠ ∅ Then
 3.    y ← FindSubstituteNNLeaf(x)
 4.    d_f(x) ← d_f(x) + d(x, y)
 5.    Copy object of y into node x
 6. Else N(b) ← N(b) − {x}

FindSubstituteNNLeaf(Node x): Node
 1.  Q ← ∅, L ← ∅
 2.  For v ∈ N(x)
 3.     If N(v) = 0 Then L ← {(v, x)}
 4.     Else Q ← {v}
 5.  While Q not empty
 6.     b ← first element of Q, Q ← Q − {b}
 7.     For v ∈ N(b)
 8.        If N(v) = 0 Then L ← L ∪ {(v, b)}
 9.        Else Q ← Q ∪ {v}
10.  (y, v) ← argmin_{(z,t)∈L} d(x, z), N(v) ← N(v) − {y}
11. Return y
```

in level-order. Finally, when the full set of leaves of the subtree of x is determined, we select the leaf y that satisfies: $d(x, y) < d(x, z), \forall (z, t) \in L - \{(y, v)\}$, then y is returned after it is disconnected from its father.

This new method is based on the idea to obtain a better deletion strategy by considering the best characteristics of GH1 and GH3: only one ghost hyperplane appears after each deletion, and its displacement is nearby to the possible best one. Clearly, we can also set a maximum allowable proportion α of ghost hyperplanes, and rebuild the tree when this limit is exceeded.

4 Experimental Results

As it is aforementioned, we do not consider the cost to locate the element as part of the deletion problem, then the deletion costs obtained represent only the necessary work to effectively delete the element from the *DSAT*. Thus, we can directly compare our experimental results with those presented in [8]. To study the behavior and performance of this new deletion algorithm for *DSAT*, we need to evaluate the proper deletion costs and also the searching performance after that.

In order to make a fairly comparison between the previous deletion methods and the new one, we use the same metric spaces considered in [8] to evaluate the performance of *DSAT*, available from [4]. We use the best arity for each space, as it is described in [8]. They are four real-life metric spaces with widely different histograms of distances: ***Strings:*** a dictionary of 69,069 English words. The distance is the *edit distance*, that is, the minimum number of character insertions, deletions and substitutions needed to make two strings equal.

NASA images: a set of 40,700 20-dimensional feature vectors, generated from images downloaded from NASA. [1] The Euclidean distance is used.
Color histograms: a set of 112,682 8-D color histograms (112-dimensional vectors) from an image database. [2] Any quadratic form can be used as a distance, so we chose Euclidean distance.
Documents: a set of 1,265 documents under the Cosine similarity, heavily used in Information Retrieval [1]. In this model the space has one coordinate per term and documents are seen as vectors in this high dimensional space. The distance we use is the angle among the vectors. The documents are the files of the TREC-3 collection.[3]

There are two types of experiments:

1. We build the index with the 90% of the database elements, the other 10% is used as queries for range searches. After the index is built, we delete a 10% of elements randomly selected.
2. We use the 60%, 70%, 80%, and 90% of the database elements to build the index. Then we delete the 10%, 20%, 30%, and 40% respectively, in order to leave 50% of the elements into the tree in each index. It can be noticed that in each case the 50% of the database, that remains after deletions into the tree, is not necessarily the same set of elements, as the elements deleted are randomly selected. Then, we perform queries with the non-inserted 10% of database elements.

We have tested several options in our experiments. For the parameter α we consider: 0% (without any ghost hyperplane), 1%, 3%, 10%, 30%, and 100% (without any rebuilding). In all cases, if $\alpha = 0\%$, as is pure rebuilding, costs are higher. Then, as the proportion α of allowed ghost hyperplanes grows, deletion costs decrease. For range search we consider three radii for the spaces with continuous distance, and four radii for Strings space (with discrete distance). For lack of space we only show some examples: for the first type of experiment, the comparison of deletion costs for $\alpha = 1\%$ when the 10% of elements is deleted; for the second one, the comparison of search costs obtained after 40% of elements is deleted with $\alpha = 1\%$, considering the 10% of reserved elements as queries. Figure 1 shows, for the first type of experiment, the average deletion costs obtained per element when 10% of the elements is deleted. Figure 2 illustrates, for the second type of experiment, the average search costs per element, after 40% of the database is deleted using $\alpha = 1\%$, when we search with the reserved 10% of elements as queries. As it can be noticed, our deletion method (GH4) obtains very good performance, both in deletion and search costs, for all metric spaces considered.

5 Conclusions

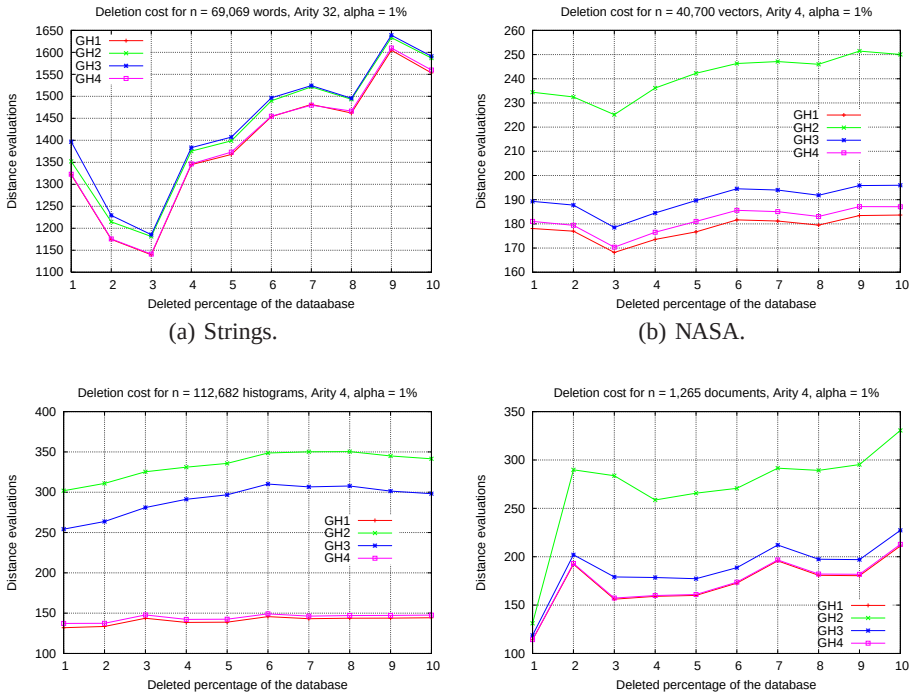
We have designed a new algorithm for efficient deletion in *DSAT*. This new algorithm has a low cost of deletion and allows that subsequent searches have a performance similar to the best algorithm proposed in [8]. On the other hand, efficient searches are still maintained: it is possible to apply the same algorithms of *DSAT* for range search and *k* closest neighbors. Our deletion algorithm kept, as a parameter, the proportion of

[1] At `http://www.dimacs.rutgers.edu/Challenges/Sixth/software.html`
[2] At `http://www.dbs.informatik.uni-muenchen.de/~seidl/DATA/histo112.112682.gz`
[3] At `http://trec.nist.gov`

allowed nodes with ghost hyperplanes in the tree, which permits us to tune search cost versus deletion cost.

The outcome is a much more practical data structure that can be useful in a wide range of applications. We expect the *DSAT*, with the new deletion algorithm, to replace the static version in the developments to come.

As future work we plan to add our new deletion algorithm to the existing version of *DSAT* for secondary memory [9], since it has the advantage that only one ghost hyperplane is created, so only two nodes have to be changed, for each deletion. In this case will be relevant both number of distance evaluations and number of I/O operations.



(a) Strings. (b) NASA.

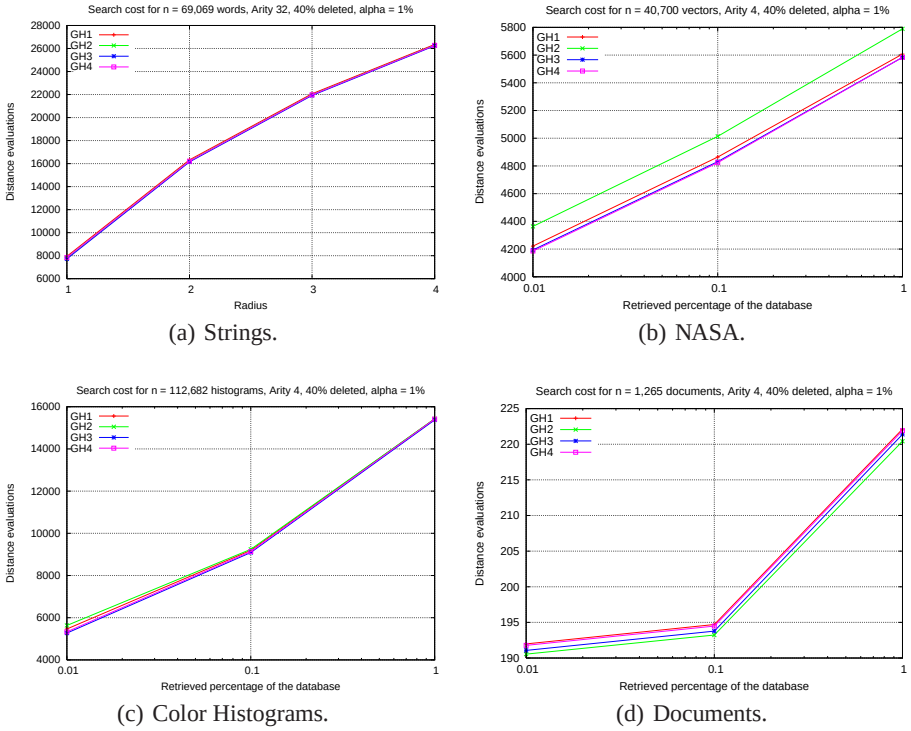
(c) Color Histograms. (d) Documents.

Fig. 1. Comparison of deletion costs, for all deletion algorithms using $\alpha = 1\%$.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
2. S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conference on Very Large Databases (VLDB'95)*, pages 574–584, 1995.
3. E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

4. K. Figueroa, G. Navarro, and E. Chávez. Metric spaces library, 2007. Available at `http://www.sisap.org/Metric_Space_Library.html`.
5. G. Hjaltason and H. Samet. Incremental similarity search in multimedia databases. Technical Report CS-TR-4199, University of Maryland, Computer Science Department, 2000.
6. G. Navarro. Searching in metric spaces by spatial approximation. In *Proc. String Processing and Information Retrieval (SPIRE'99)*, pages 141–148. IEEE CS Press, 1999.
7. G. Navarro. Searching in metric spaces by spatial approximation. *The Very Large Databases Journal (VLDBJ)*, 11(1):28–46, 2002.
8. G. Navarro and N. Reyes. Dynamic spatial approximation trees. *ACM Journal of Experimental Algorithmics (JEA)*, 12:article 1.5, 2008. 68 pages.
9. G. Navarro and N. Reyes. Dynamic spatial approximation trees for massive data. In T. Skopal and P. Zezula, editors, *SISAP*, pages 81–88. IEEE Computer Society, 2009.
10. R. Uribe and G. Navarro. Una estructura dinámica para búsqueda en espacios métricos. In *Actas del XI Encuentro Chileno de Computación, Jornadas Chilenas de Computación*, Chillán, Chile, 2003. In Spanish. In CD-ROM.



(a) Strings. (b) NASA.

(c) Color Histograms. (d) Documents.

Fig. 2. Comparison of search costs, after 40% of elements are deleted using $\alpha = 1\%$.

Prototipo de búsqueda y comparación que aplica técnicas de recuperación de información en bases de datos relacionales

Claudio Camacho, Walter Singer y Rosanna Costaguta

Departamento de Informática, Facultad de Ciencias Exactas y Tecnologías (FCEyT)
Universidad Nacional de Santiago del Estero (UNSE)
Avda. Belgrano (S) 1912, Santiago del Estero, 4200, Argentina
claudiocamacho@yahoo.com, singerwalter@gmail.com, rosanna@unse.edu.ar

Resumen. Una de las actividades más importante de una organización comercial es la adquisición de productos, ya que el encargado de compras debe tener en cuenta diversos aspectos como precios, ofertas, disponibilidad, y tiempo de entrega, entre otras cuestiones, a fin de tomar una decisión. Esto se vuelve una tarea complicada cuando la organización posee información de numerosos artículos que provienen de distintos proveedores, donde cada uno de ellos obviamente administra sus propias listas de precios y catálogos de productos, con nomenclaturas diferentes (código de artículo, descripción, etc.). Es así que al momento de, por ejemplo, comparar precios entre los distintos proveedores, no es fácil identificar productos equivalentes. Debido a lo expuesto resulta conveniente que el Sistema de Información de la organización posea ciertas características de búsqueda que permitan responder eficientemente a estas cuestiones. En este artículo se presenta un prototipo de herramienta de búsqueda aplicable sobre bases de datos relacionales, que utiliza técnicas de recuperación inteligente de información. Las pruebas realizadas arrojaron resultados muy satisfactorios.

Palabras clave: Técnicas de Recuperación Inteligente de Información, Bases de Datos Relacionales, Sistemas de Información, Organizaciones Comerciales

1 Introducción

En la actualidad es cada vez mayor la utilización de Sistemas de Información (SI) en las distintas áreas de las organizaciones comerciales, así como también el incremento en los volúmenes de datos almacenados en sus Bases de Datos (BD). Debido a esto, la obtención de información a través de un determinado proceso se hace cada vez más ineficiente ya que el tiempo de respuesta de las consultas sobre los datos, es cada vez mayor y con resultados menos precisos.

En particular en las organizaciones comerciales, en el área de compras es donde se concentran las decisiones que se deben tomar con respecto a la mejor opción de compra, y es por ello que generalmente se deben comparar los precios de los artículos en las listas de precios de los diferentes proveedores que comercializan un mismo

producto. Realizar este proceso en forma manual requiere mucho tiempo, ya que cada artículo posee una nomenclatura propia (código, descripción, marca, etc.) para cada uno de los proveedores, y es aquí donde surge la necesidad de utilizar métodos y técnicas de búsqueda y clasificación más eficientes que ayuden a los usuarios a identificar artículos equivalentes o iguales.

Dado lo expuesto, en este artículo se presenta una herramienta de RI y clasificación de los resultados para el SI de una organización comercial del medio, cuyo uso permitió optimizar los procesos de búsqueda y comparación por parte del usuario. El prototipo desarrollado utiliza librerías de RI (Sphinx 0.9.9-rc2) y técnicas de clasificación. La presente investigación fue desarrollada como trabajo final de graduación de dos de los autores, a fin de obtener el título de Licenciado en Sistemas de Información.

Este artículo se organiza como sigue. En la sección 2 se describe brevemente la problemática que dio origen a este trabajo. La sección 3 contiene antecedentes relevantes. La sección 4 describe el prototipo desarrollado. La sección 5 documenta las pruebas efectuadas y el análisis de los resultados obtenidos. La sección 6 enuncia algunas conclusiones sobre el trabajo realizado.

2 Planteamiento del problema

Son muchas las tareas que se realizan en una organización comercial, una de las más importantes es la que involucra al área de compras. En esta área el encargado debe analizar y comparar artículos en extensas listas de artículos de distintos proveedores teniendo en cuenta ciertos factores como precio, marcas, tiempo de entrega de la mercadería, disponibilidad en stock en los proveedores, entre otros. Esta es una tarea que suele consumir un tiempo considerable sobre todo en el momento de identificar los artículos. Esta labor se vuelve aun más compleja si se tiene en cuenta que la nomenclatura (códigos, descripciones, marca, etc.) que se utiliza varía de un proveedor a otro. Existen casos en los que, por ejemplo, se suprime parte del código del artículo, en otros casos se agrega un código complementario al código del artículo, y hasta en otros casos el Código del artículo se sustituye completamente y el mismo es agregado como parte de la descripción del artículo. Así, puede ocurrir que un mismo artículo se presente dentro de la BD de maneras diferentes, con lo que la tarea de filtrar estos datos a través de una consulta estructurada es casi imposible dada la cantidad de variantes con la cuales se puede identificar un artículo. En la Tabla 1 puede observarse un ejemplo extraído del catálogo real de Cables Marca *"NGK"*. El artículo *"CABLE DE BUJIA VW GOL*", cuyo código es *"ST-V02",* aparece cargado tres veces en la tabla artículos de la organización ya que es suministrado por tres proveedores diferentes. Sin embargo, como puede observarse, el código de artículo ST-V02 aparece cargado como código de artículo en el caso del primer proveedor, como parte del código de artículo en el caso del segundo proveedor y como parte de la descripción del artículo para el tercer proveedor. Así, como puede observarse en el ejemplo de la Tabla 1, un mismo artículo se presenta dentro de la BD de la organización de maneras diferentes, y por tanto, la tarea de filtrar estos datos a través

de una consulta estructurada es casi imposible por la cantidad de variantes con la cuales se puede identificar un artículo dado.

Tabla 1. Fragmento de la BD de la organización

Código	Descripción de Artículo	Cód. Proveedor
STV02	CABLE DE BUJIA	1
…	…	…
0605STV02	CABLE BUJIA SEN/ GOL	5
…	…	…
NGK-02	CABLE DE BUJIA NGK ST V02	12
…	…	…

3 Antecedentes relacionados

A continuación se describen brevemente dos trabajos relevantes que integran sistemas de recuperación de información con sistemas de bases de datos estructuradas o semiestructuradas. Sin embargo, cabe destacar que no se encontraron antecedentes en el ámbito comercial como el que se desarrolla aquí.

En el ámbito de la salud se encontró un proyecto aplicado sobre la BD biomedicinal MEDLINE [1]. Esta cuenta con una BD donde cada registro almacena la referencia bibliográfica de un artículo científico publicado en una revista médica, conteniendo además datos básicos como título, autores, etc., posibilitando su recuperación a través de Internet. Los autores desarrollaron dos sistemas de indexación y búsqueda utilizando dos tecnologías aplicadas al tratamiento de los datos (LUCENE[1] y PostgreSQL[2]) que mejoró la capacidad de búsqueda y recuperación de información en MEDLINE. Una vez construidos ambos sistemas de búsqueda, éstos fueron evaluados en cuanto a rendimiento para luego decidir cuál era el más apropiado para manejar la base de datos de MEDLINE. Los autores eligieron a LUCENE porque está optimizada para bases de datos textuales y por tanto ofrece mejores posibilidades para tratar datos no estructurados.

En [2] se extiende la estructura de un Sistema de Gestión de Base de Datos (SGBD) semiestructurada (XML) para agregar funciones de recuperación de información a las consultas estándares. La investigación se dividió en dos etapas: Diseño de la Arquitectura de la base de datos y Optimización de las consultas. Se formuló un modelo dividido en tres capas: Física, Lógica y Conceptual. La primera etapa de desarrollo abarcó las dos primeras capas, definiéndose un pequeño número de primitivas de recuperación en la capa lógica del SGBD, como una extensión de la arquitectura actual, para proveer consultas que combinen ambos enfoques (estructurado y no estructurado). Para evitar la redundancia en el procesamiento de las consultas los autores proponen crear reglas de optimización asociadas a las primitivas

[1] Lucene es un grupo de librerías escritas en JAVA que brinda las primitivas necesarias para el tratamiento de datos textuales en la recuperación de información.

[2] PostgreSQL es un SGBD similar a MySQL u Oracle, pero más robusto que estos dos últimos cuando se necesita operar con BD grandes

de estas consultas combinadas. Esta tarea es la que se realizará en la segunda etapa. Como resultado [3] se obtuvo un prototipo que realiza búsquedas en la colección de documentos de la base de datos y en la estructura de la misma (etiquetas XML) para obtener así información más certera.

4 El prototipo desarrollado

El prototipo se desarrolló teniendo en cuenta dos procesos principales: Proceso de RI y Proceso de Clasificación, los mismos a su vez están divididos en subprocesos. El proceso de RI está compuesto por los subprocesos: *Depuración de datos, Indexación de datos y Búsqueda*. El proceso de clasificación está compuesto por los subprocesos: *Entrenamiento del clasificador y Clasificación de los resultados*. Cada uno de estos subprocesos contempla la utilización de diferentes algoritmos y librerías de RI, así como también lenguajes de programación como SQL y Delphi. El motor de BD que se utilizó fue MySQL Server 5.0.

4.1. Los subprocesos de RI

La *depuración de datos* implicó realizar tareas de limpieza sobre la BD, como ser la eliminación de caracteres extraños (por ej. &, %, #, ã, Ø, ©, Œ) y espacios innecesarios. Este subproceso de depuración es necesario para optimizar los resultados de la indexación. La depuración se realiza a nivel de capa de datos, es decir, como procedimientos almacenados en la BD y convocados por el prototipo, pero ejecutados por el motor de la BD. Cada vez que se ingresan nuevos registros o se modifican registros existentes, el prototipo ejecuta este subproceso para depurar solamente estos registros.

La *indexación* genera un conjunto de archivos externos a la BD, cuya estructura es propia de la herramienta de RI utilizada (Sphinx). Este subproceso se realiza siempre que se agreguen nuevos datos a la BD. El prototipo, accede a los datos a través de una consulta sobre la tabla de interés, en la cual se seleccionan los campos que se desean indexar.

El *subproceso de búsqueda* es el encargado de procesar las consultas de los usuarios utilizando el índice generado por el subproceso de Indexación. El prototipo convoca al subproceso de búsqueda para realizar una búsqueda, así, a partir de una consulta de usuario utiliza los archivos índice y devuelve un listado de artículos relevantes. Luego genera una tabla índice dentro de la BD con el campo identificador de la tabla de datos, seguidamente realiza una unión de la tabla índice con la tabla de datos, y de esta manera puede mostrar información más detallada de los registros encontrados como resultado final de la consulta.

4.2. Los subprocesos de Clasificación

El subproceso de clasificación utiliza una modificación de los algoritmos propuestos por [4]. Aquí se clasifica cada registro de la lista obtenida por el subproceso de búsqueda teniendo en cuenta una categoría en particular. En primera instancia, se deduce o clasifica la consulta ingresada por el usuario para obtener cual es la

categoría más probable a la que pertenece. En segunda instancia, el clasificador utiliza dicha categoría para clasificar la lista de registros. El criterio para ambos pasos es similar. Así el subproceso de clasificación se separa en dos subprocesos, en el primero se clasifica la consulta del usuario y se obtiene la categoría más probable, y en el segundo subproceso se ordenan los registros encontrados mediante la búsqueda y utilizando la categoría encontrada en el proceso anterior. Durante el entrenamiento del clasificador se utilizó el algoritmo de Naives Bayes [4]. A continuación se presenta el modelo bayesiano utilizado en el subproceso.

Sean C un conjunto de clases tal que C={P,NP} donde P="Pertenece" y NP="No Pertenece", C'={c_1',c_2',…,c_m'} un conjunto de categorías definidas por el usuario en base a algún criterio subjetivo en donde cada una de éstas categorías toma los estados P y NP, X un conjunto de todos los artículos de la BD de la organización tal que $X=\{x_1,x_2,...,x_n\}$ y $V=\{t_1,t_2,...,t_k\}$ un conjunto de términos seleccionados en el proceso de indexación de la BD, se define la red bayesiana mostrada en la Figura 1 para realizar la clasificación de una lista de artículos en donde C_i representa las categorías definidas por el usuario y t_1, …, t_k son los términos claves que fueron indexados en el proceso de indexación.

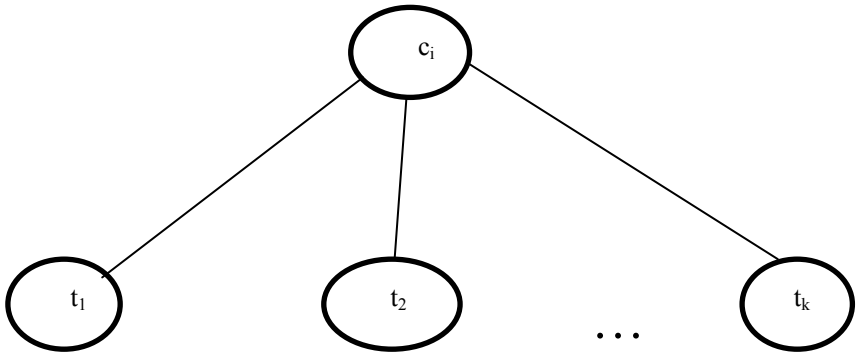


Fig. 1. Red bayesiana para la clasificación de listado de registros

Teniendo en cuenta la ecuación (1), la probabilidad a priori de c_i ($P(c_i)$) y las probabilidades a posteriori de t_j ($P(t_j/c_i)$) son estimadas en el proceso de entrenamiento sobre un conjunto inicial de artículos (registros). Para cada nuevo registro el prototipo convoca al subproceso de clasificación para la clase *P* y *NP*, luego obtiene el máximo valor de probabilidad entre P(*P*) y P(*NP*) para determinar si pertenece o no a una categoría $c_k^{'}$ determinada.

$$c_{MAP} = \arg\max_{c\in\square} P(c \mid d) = \arg\max_{c\in\square} P(c) \prod_{1\le k\le n_d} P(t_k \mid c) \qquad (1)$$

En el subproceso de entrenamiento, los parámetros que se deben tener en cuenta son: las categorías elegidas por el usuario, el conjunto de registros de entrenamiento y las dos clases, Pertenece (P) y No Pertenece (NP). Cada categoría posee estas dos clases. El algoritmo irá determinando las probabilidades P y NP para cada palabra de un conjunto de entrenamiento seleccionado a priori, considerando cada categoría

definida. Cabe aclarar que estas palabras se obtuvieron mediante un algoritmo de selección de términos relevantes.

Cuando se preparan los datos para el entrenamiento debe tenerse en cuenta que existe la posibilidad de que el usuario pueda ingresar nuevas categorías en cualquier momento. Además, luego de una clasificación de los resultados, el usuario podría clasificar manualmente cada registro para mejorar la exactitud del clasificador.

El subproceso de clasificación identifica cual es la categoría más probable a la que pertenece la consulta, para esto se utiliza un algoritmo que permite clasificar e identificar entre todas las categorías cual es la más probable. El algoritmo devuelve un conjunto de registros *L*. Para clasificar la lista de resultados *L* devueltos por el buscador, se convoca a rutinas que permiten obtener un coeficiente de ordenamiento para cada registro de la lista. Este coeficiente indica la relevancia que tiene un registro con respecto a la consulta del usuario. Una vez obtenido el conjunto de coeficientes asociados a cada registro del listado *L*, éste es ordenado en forma ascendente utilizando dicho coeficiente. Se obtiene así un listado de registros ordenados por relevancia. La Tabla 2 muestra los procesos y subprocesos que componen el prototipo desarrollado.

Tabla 2. Componentes del prototipo desarrollado

Procesos	Subprocesos	Descripción
Proceso de RI	Depuración	Elimina caracteres extraños.
	Indexación	Busca en la BD palabras claves y los indexa guardando en un archivo externo.
	Búsqueda	Utiliza el archivo índice generado para realizar búsqueda además de operaciones relacionales para obtener información más detallada sobre cada registro (artículos).
Proceso de Clasificación	Entrenamiento	Se realiza con un conjunto inicial de registros previamente clasificados por el usuario, de manera encontrar una función que permita clasificar registros cuya categoría se desconozca.
	Clasificación de los resultados	Determina a que categoría pertenecen nuevos registros de la lista de resultados devueltos por el proceso de búsqueda.

5 Experimentación y análisis de resultados

Las pruebas se realizaron considerando tres escenarios diferentes (*Sistema Actual, Prototipo de búsqueda sin clasificador y Prototipo de búsqueda con clasificador*). En los mismos se utilizaron como casos de prueba diez consultas especialmente formuladas por el encargado de compras de la organización, por ser ejemplos de búsquedas típicas en situaciones reales. Para evaluar los resultados obtenidos para los casos de prueba en cada uno de los tres escenarios propuestos se definieron cinco indicadores de rendimiento. La Tabla 3 muestra la unidad de medida y una pequeña descripción de cada uno de los indicadores utilizados.

Tabla 3. Indicadores de rendimiento utilizados

Indicadores	U. de medida	Descripción
Tiempo de respuesta (TR)	Milisegundos	Tiempo que el usuario debe esperar antes de obtener los resultados de su consulta
Cantidad de resultados (CR)	Registros	Cantidad de resultados devueltos por el sistema
Velocidad de respuesta (VR)	Registros por milisegundo	Cantidad de registros por milisegundo que devuelve el sistema
Precisión (P)	Valor real entre 0 y 1	Proporción de registros relevantes dentro del conjunto de registros recuperados
Exhaustividad (E)	Valor real entre 0 y 1	Proporción de registros relevantes de la BD que fueron recuperados

A fin de facilitar la comparación de los resultados de las pruebas realizadas con el SI actual (escenario 1), con el prototipo de búsqueda desarrollado pero sin el clasificador (escenario 2) y con el prototipo de búsqueda incluyendo el clasificador (escenario 3), los mismos se resumen en la Tabla 4.

Como puede observarse, el indicador *Tiempo de respuesta (TR*) es el que más evidencia aporta para demostrar que el uso del prototipo es más eficiente. Utilizando el prototipo sin clasificador los tiempos de búsqueda se reducen hasta alcanzar una diferencia de velocidad media (Vm) 120 veces mayor a la del sistema actual, y cuando se utiliza el prototipo con clasificador la diferencia de velocidad media es aproximadamente 11 veces mayor a la del sistema actual de la organización. Si bien en los dos escenarios que se utilizó el prototipo de búsqueda TR disminuyó considerablemente, al utilizar el clasificador el TR es mayor por cuanto el prototipo debe realizar operaciones adicionales para cumplir con el objetivo de clasificar la lista de resultados antes de ser devuelta al usuario. Con respecto al indicador Cantidad de resultados (CR) puede afirmarse que no se observan cambios significativos entre los valores arrojados en cada escenario, existiendo sólo una variación de 30 registros en promedio.

Comparando el indicador *Velocidad de respuesta (VR)* se observa que el prototipo de búsqueda sin clasificador es en promedio 13 veces más rápido que el prototipo con clasificador. Esto resulta lógico ya que como se explicó el prototipo de búsqueda que incluye al clasificador debe realizar tareas adicionales para clasificar los resultados antes de mostrárselos al usuario. La diferencia de velocidad media se calcula utilizando la ecuación (2):

$$Vm = \left(\sum_{1}^{n} \left({tr1_i}/{tr2_i} \right) \right)/n \qquad (2)$$

Siendo *tr1* y *tr2* los tiempos de respuesta del sistema actual y del prototipo respectivamente, y *n* la cantidad de pruebas realizadas. El resultado obtenido de la división entre tr1 y tr2 indica cuantas veces más rápido es el prototipo de búsqueda que el sistema actual. Si $tr1/tr2 > 1$ entonces el prototipo es más rápido que el SI, si $tr1/tr2 = 1$ indica que el prototipo y el SI tienen el mismo TR, si $tr1/tr2 < 1$, el prototipo es más lento que el SI actual. En la ecuación (2) se busca obtener un valor

Vm que indicará cuanto más rápido es el prototipo de búsqueda que el SI actual y debido a esto se calcula el promedio de los resultados de las divisiones de los TR.

Tabla 4. Comparación de los indicadores obtenidos en los tres escenarios de prueba

Casos de prueba	Escenarios	Indicadores de rendimiento				
		TR	**CR**	**VR**	**P**	**E**
amort* bora	SI actual	15702	57	0,004	0,614	1
	Prototipo de RI	140	45	0,321	0,511	0,657
	Prototipo de RI con Clasif.	2780	45	0,016	0,511	0,657
rotula golf	SI actual	19030	52	0,003	0,731	1
	Prototipo de RI	172	32	0,186	0,813	0,684
	Prototipo de RI con Clasif.	2125	32	0,015	0,813	0,684
bujia corsa	SI actual	18186	52	0,003	0,500	1
	Prototipo de RI	62	21	0,339	0,714	0,577
	Prototipo de RI con Clasif.	1281	21	0,016	0,714	0,577
cable bujia gol*	SI actual	19905	99	0,005	0,909	1
	Prototipo de RI	203	36	0,177	1,000	0,400
	Prototipo de RI con Clasif.	1906	36	0,019	1,000	0,400
correa 6pk* gol	SI actual	19530	86	0,004	1,000	1
	Prototipo de RI	235	39	0,166	1,000	0,453
	Prototipo de RI con Clasif.	2390	39	0,016	1,000	0,453
rot* gol thompson	SI actual	19811	52	0,003	1,000	1
	Prototipo de RI	79	28	0,354	1,000	0,538
	Prototipo de RI con Clasif.	1468	28	0,019	1,000	0,538
correa dist* gol*	SI actual	19201	110	0,006	0,536	1
	Prototipo de RI	250	78	0,312	0,372	0,492
	Prototipo de RI con Clasif.	3999	78	0,020	0,372	0,492
bujia golf	SI actual	15218	62	0,004	0,468	1
	Prototipo de RI	188	24	0,128	0,708	0,586
	Prototipo de RI con Clasif.	1718	24	0,014	0,708	0,586
filtro aire gol*	SI actual	1998	79	0,040	0,709	1
	Prototipo de RI	297	65	0,219	0,677	0,786
	Prototipo de RI con Clasif.	3499	65	0,019	0,677	0,786
optica corsa vic	SI actual	20076	32	0,002	0,688	1
	Prototipo de RI	234	13	0,056	0,538	0,318
	Prototipo de RI con Clasif.	781	13	0,017	0,538	0,318

Respecto al indicador *Precisión (P)* se observa que en dos casos de prueba no existen diferencias en los resultados obtenidos usando el SI actual y el prototipo de

búsqueda (con y sin clasificador), en cuatro casos el uso del prototipo mejora los resultados pero en otros cuatro los empeora. Dada esta situación se decidió calcular un valor promedio para el indicador, obteniéndose 0,715 para las pruebas con el SI actual y 0,733 con el prototipo. Estos resultados son un indicio de que el prototipo de búsqueda desarrollado es un poco más eficiente que el SI actual en la obtención de resultados relevantes.

En cuanto al indicador *Exhaustividad (E)*, se observó que el SI actual tiene mejor desempeño. Esto se debe en gran medida a la forma en la que se realizan las búsquedas en dicho sistema. Cabe destacar que este indicador por sí solo no demuestra nada, ya que si bien los valores llegan al máximo, se tiene que analizar la cantidad total de registros devueltos por la consulta y el tiempo de respuesta de la misma. En particular, los resultados arrojados por el prototipo de búsqueda, con y sin clasificador, presentan un valor medio es de 0,61 que se ubica por encima de la media general del indicador. Este valor medio podría mejorar ya que dependen en gran medida del conocimiento y practica que adquiera el usuario en la formulación de las consultas, es decir, cuanto más precisa sea la consulta del usuario, mejores resultados devolverá el prototipo, ya sea por la experiencia del mismo o por el uso de atajos que brinda el prototipo como son los comodines.

6 Conclusiones

El desarrollo del prototipo permitió al SI de la organización comercial incrementar su capacidad de búsqueda, ya que el mismo fue integrado satisfactoriamente a dicho SI. Luego del análisis de las pruebas realizadas al SI y al prototipo de búsqueda se pudo observar que se han mejorado notablemente los tiempos de respuesta, así como también la calidad de los resultados de las consultas de usuario. Esto se debe a que con la herramienta de RI utilizada (Sphinx) y la aplicación de técnicas de clasificación sobre los resultados, se logró obtener una lista refinada y ordenada de resultados considerando su relevancia respecto a la consulta de usuario. Además, es de resaltar que la precisión en los primeros puestos del ranking mejora a medida que el usuario realiza las consultas y el entrenamiento continuo del prototipo.

Con el prototipo de RI desarrollado la organización comercial objeto de estudio logró obtener familias de artículos semánticamente similares en los cuales se encontraron registros sintácticamente diferentes, es decir, el prototipo devolvió artículos relevantes de diferentes proveedores cuyos códigos poseen diferente nomenclatura. Además, se pudo comprobar que incorporando tecnologías de la información como las técnicas de Minería de Datos y la Inteligencia Artificial a las técnicas convencionales, se resuelven estas situaciones de manera más eficiente que utilizando solamente procedimientos, técnicas y métodos convencionales.

Actualmente se está considerando aplicar otras técnicas de clasificación, como por ejemplo, el vecino más cercano [4], a fin de comprobar si es posible mejorar aún más el desempeño del prototipo desarrollado.

Referencias

1. García, F., Fernández, Ch., Azancot, M. (2007). Desarrollo de un sistema de indexación y búsqueda sobre la base de datos de biomedicina MEDLINE. Recuperado el 13 de marzo de 2013 en http://biblioteca. universia.net/html_bura/ ficha/params/id/45165231.html
2. Hiemstra, D., Vries, A., Blok, H., Keulen, M., Jonker, W., Kersten, M. CIRQUID: Complex Information Retrieval queries in a Database. Recuperado el 13 de marzo de 2013 en http://doc.utwente.nl/47223/ 1/hiemstra03cirquid.pdf
3. Johan, L., Vojkan, M., Ramirez, G., de Vries, A., Hiemstra, D., Blok, H. (2005). "*TIJAH: Embracing IR Methods in XML Databases*", Information Retrieval Journal 8, Kluwer Academic Publishers, ISSN 1386-4564, pp. 547-570. Recuperado el 13 de marzo de 2013 en http://www.cs.utwente.nl/~hiemstra/papers/irj05.pdf
4. Manning, C., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.
5. Tolosa, G, Bordignon, F. (2008). Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos. (Universidad Nacional de Luján, Buenos Aires). Recuperado el 13 de marzo de 2013 en http://www.tyr.unlu.edu.ar/tallerIR/2008/docs/Introduccion-RI-v9f.pdf
6. Brisaboa, N., Farina, A., Pedreira, O., Reyes, N. (2007). Indexación dinámica para la recuperación de información basada en búsqueda por similitud. Recuperado el 13 de marzo de 2013 en http://www.sistedes.es/sistedes/ pdf/2007/JISBD-07-brisaboa-indexacion.pdf
7. Hernandez Orallo, J., Ramirez Quintana, M., Ferri Ramirez, C. (2004). Introducción a la Minería de Datos. Madrid, España: Prince Hall.
8. Artayer, L. (2006). Construcción de un prototipo de un Sistema de Información Basado en Ontología Trabajo final para optar por el título de Licenciado en Sistemas de Información. Universidad Nacional de Santiago del Estero

VIII WORKSHOP ARQUITECTURA, REDES Y SISTEMAS OPERATIVOS

- WARSO -

CACIC2013

VIII WORKSHOP ARQUITECTURA, REDES Y SISTEMAS OPERATIVOS

- WARSO -

ID	Trabajo	Autores
5600	**Análisis de las prestaciones de 802.11e en redes MANET**	María Antonia Murazzo (UNSJ), Nelson R. Rodriguez (UNSJ), Daniela A. Villafañe (UNSJ)
5777	**Caso de estudio de comunicaciones seguras sobre redes móviles ad hoc**	Sergio H. Rocabado Moreno (UNSa), Daniel Arias Figueroa (UNSa), Ernesto Sánchez (UNSa), Javier Díaz (UNLP)
5648	**Estudio del desempeño de OLSR en una red mallada inalámbrica en un escenario real**	Eduardo Rodríguez (UCA), Claudia Deco (UCA), Luciana Burzacca (UCA), Mauro Petinari (UCA)
5652	**NetworkDCQ: A Multi - platform Networking Framework For Mobile Applications**	Federico Cristina (UNLP), Sebastián Dapoto (UNLP), Fernando G. Tinetti (UNLP), Pablo Thomas (UNLP), Patricia Pesado (UNLP)
5669	**Estimación de "H" con transformada ondita**	Luis Marrone (UNLP), Reinaldo Scappini (UTN-FRRe)
5752	**IP Core Para Redes de Petri con Tiempo**	Orlando Micolini (UNC), Julian Nonino (UNC), Carlos Renzo Pisetta (UNC)
5836	**Analysis of Radio Communication Solutions in Small and Isolated Communities under the IEEE 802.22 Standard**	Alejandro Arroyo Arzubi (EST-IESE), Antonio Castro Letchtaler (IESE), Antonio Foti (UTN), J. Garcia Guibout (UNCUYO), Rubén Jorge Fusario (UBA), L. Sens (UTN)
5834	**Posicionamiento indoor determinado por la distancia en función de la potencia medida de balizas bluetooth**	Marcelo Marinelli (UNaM), Juan Toloza (UNCPBA), Nelson Acosta (UNCPBA)

VIII WORKSHOP ARQUITECTURA, REDES Y SISTEMAS OPERATIVOS

- WARSO -

ID	Trabajo	Autores
5835	**Posicionamiento WIFI con variaciones de Fingerprint**	Carlos Kornuta (UNCPBA), Nelson Acosta (UNCPBA), Juan Toloza (UNCPBA)
5693	**Monitoreo remoto de sistemas y redes para la auditoria informática**	María Elena Ciolli (IUA), Claudio Porchietto (IUA), Roberto Rossi (IUA), Juan Sapolski (IUA)
5863	**DJBot: Administrando las salas de PC evitando la consola**	Javier Díaz (UNLP), Aldo Vizcaino (UNLP), Alejandro Sabolansky (UNLP), Einar Felipe Lanfranco (UNLP)

Análisis de las prestaciones de 802.11e en redes MANET

María Murazzo[1*], Nelson Rodríguez[2*], Daniela Villafañe[3*]

[1]marite@unsj-cuim.edu.ar, [2]nelson@iinfo.unsj.edu.ar, [3]villafañe.unsj@gmail.com
* *Docentes e Investigadores, Departamento e Instituto de Informática – FCEFyN - UNSJ*

Abstract. El desarrollo de las redes de comunicaciones móviles y sus servicios, ha supuesto un gran esfuerzo científico y técnico para dotar de mecanismos capaces de garantizar QoS (Quality of Service)[1] a los usuarios.

En el caso de las redes móviles ad-hoc (MANET), este esfuerzo es relevante debido a la complejidad y dinamicidad del entorno. Por lo cual, para proporcionar QoS en estas redes, es importante garantizar los requerimientos necesarios y gestionar eficientemente los recursos disponibles.

Con el objeto de proporcionar la calidad demandada por las actuales aplicaciones, han surgido propuestas que abordan la problemática de la QoS en diferentes capas.

Este trabajo aborda la problemática de la provisión de QoS en redes MANET desde la perspectiva de la subcapa MAC, mediante la implementación de 802.11e, para analizar el retardo y la sobrecarga que sufren las transmisiones en ambientes con baja y alta granularidad de nodos.

Keywords: MANET, QoS, 802.11e, NS2, CBR

1 Introducción

Una red móvil ad hoc (MANET) es una red de comunicaciones formada espontáneamente por un conjunto de dispositivos móviles inalámbricos capaces de comunicarse entre sí, sin la necesidad de una infraestructura de red fija o gestión administrativa centralizada.

Estas redes nacen bajo el concepto de autonomía e independencia, al no requerir el uso de infraestructura pre-existente ni una administración centralizada como las redes cableadas.

Debido a que el alcance de transmisión de los dispositivos es limitado, pueden llegar a ser necesarios nodos intermedios para transferir datos de un nodo a otro. Por ello, en una red MANET cada nodo puede operar como fuente, destino o router (naturaleza “multihop”).

[1] En español, calidad de servicio.

En estas redes, los nodos son libres para moverse arbitrariamente, produciendo cambios en la topología de la red. El grado de movilidad y cambio de la topología depende de las características de los nodos. Las variaciones en el canal de radio y las limitaciones de energía de los nodos pueden producir cambios en la topología y en la conectividad.Por lo que, las MANET deben adaptarse dinámicamente para ser capaces de mantener las conexiones activas a pesar de estos cambios [1].

2 QoS en Redes MANET

Las actuales redes de telecomunicación, y principalmente las MANET, se caracterizan por un constante incremento del número, complejidad y heterogeneidad de los recursos que las componen. Esta heterogeneidad, se pone aun más de manifiesto según el tipo de aplicaciones que se ejecutan. En la actualidad, la mayoría del tráfico de red es generado por aplicaciones de tipo multimedial o con fuertes restricciones en cuanto a la cantidad de recursos que demandan de la red.

El tráfico multimedia, como el utilizado en telefonía IP o videoconferencia, puede ser extremadamente sensible a los retardos y puede crear demandas de QoS muy restrictivas sobre las redes que los transportan. Cuando los paquetes son entregados usando el modelo de mejor esfuerzo, estos no arriban en una manera oportuna. El resultado son imágenes no claras, desiguales, movimientos lentos, y el sonido no se lo obtiene sincronizado con la imagen.

Los aspectos críticos que causan la mayor parte de los problemas en aplicaciones con grandes restricciones de recursos son: *falta de ancho de banda*, *retardo extremo a extremo* y *pérdida de paquetes.*

Respecto a la falta de ancho de banda, la mejor opción para contrarrestarlo, es clasificar el tráfico dentro de clases de QoS y priorizar tráfico de acuerdo a la importancia del mismo.

El retardo extremo a extremo, se puede disminuir dándole a los paquetes pertenecientes a aplicaciones sensitivas, cierta prioridad para que en el camino extremo a extremo sean tratados de manera más ágil.

Por último, y en relación a la perdida de paquetes, estos pueden ser descartados cuando un enlace está congestionado. Un paliativo para esta problemática es un esquema de scheduler de tráfico que permita proporcionar un mejor servicio a paquetes pertenecientes a aplicaciones sensibles [2].

La QoS, es un término usado para definir la capacidad de una red para proveer diferentes niveles de servicio a los distintos tipos de tráfico. Permite que los administradores de una red puedan asignar a un determinado tráfico prioridad sobre otro y, de esta forma, garantizar que un mínimo nivel de servicio le será provisto.

Aplicando técnicas de QoS se puede proveer un servicio más acorde al tipo de tráfico y de esta manera permitir: priorizar ciertas aplicaciones de nivel crítico en la red, maximizar el uso de la infraestructura de la red, proveer una mejor performance a aplicaciones sensitivas al delay como son las de voz y video y responder a cambios en los flujos del tráfico de red.

Al aplicar técnicas de QoS, el administrador de la red puede tener control sobre los diferentes parámetros que definen las características de un tráfico en particular (retardo extremo a extremo, latencia, variaciones de latencia, perdida de paquetes, ancho de banda).

El problema de la administración de QoS, está prácticamente resuelto en redes fijas, pero esto no sucede en redes inalámbricas y específicamente en redes MANET cuyas características hacen necesario un nuevo estudio para afrontar este problema.

La topología dinámica, la naturaleza multihop y los escasos recursos de los nodos hacen necesario que los mecanismos de provisión de QoS sean lo más ligeros posibles, en cuanto a carga de procesamiento, como de recursos de red (ancho de banda), para evitar que el throughput o capacidad disponible por nodo se reduzca drásticamente [3].

De todo lo enunciado se puede llegar a la conclusión que la única manera de poder lograr la coexistencia de aplicaciones con diferentes niveles de requerimientos de recursos es realizando una adecuada administración del trafico mediante la priorización implícita o explícita de los paquetes de datos.

La priorización de tráfico, permitirá que ciertos flujos de datos puedan ser tratados de forma preferencial logrando maximizar el uso del ancho de banda, minimizar el retardo extremo a extremo y minimizar la perdida de paquetes. Esta priorización se logra mediante la implementación de mecanismos de QoS que permita una gestión de los flujos de tráfico [4].

Existen diversas formas de proveer mecanismos de QoS a las redes: ruteo con QoS, señalización de QoS y MAC (Medium Access Control) con QoS. De todas estas opciones en este trabajo se seleccionó el análisis del impacto de la implementación de mecanismos de QoS en la capa MAC mediante el uso del estándar 802.11e.

2.1 QoS en Capa MAC

Las características de movilidad de los nodos que pertenecen a una red MANET hacen muy difícil la administración de QoS en los niveles de red. Dicha movilidad produce una sobrecarga excesiva, lo cual produce un empeoramiento en la eficiencia de la red.

El estándar 802.11e no congestiona la red con paquetes de señalización, ni con paquetes de descubrimiento de rutas, sino que plantea una forma de administración general, la cual se basa en los tiempos de espera. El que tiene mayor prioridad de transmisión es el que menos tiempo debe esperar. Esto hace, que no sea necesario inundar la red con paquetes, pues cada nodo por separado sabrá si tiene que transmitir o no de acuerdo al tiempo que tenga que esperar, de este modo cada nodo transmite de forma independiente, lo cual evita la necesidad de sincronizarse con los demás nodos para reservar recursos y asignar prioridades de forma conjunta. Otra diferencia importante es que se provee QoS a los paquetes o flujos de datos y no a los nodos que están transmitiendo, a diferencia del ruteo QoS el cual se encarga de hacer reserva de recursos de acuerdo a las capacidades de los nodos; por esta razón con 802.11e no hay necesidad de sincronizarse con los nodos de la ruta seleccionada. Por otra parte, al ser

tan dinámica, las MANET tienen caídas de enlaces muy frecuentemente, lo cual hace que se deban realizar procesos de retransmisión. Dichos procesos en el estándar 802.11e se realizan ejecutando algoritmos de backoff sin la necesidad de reenvíos de paquetes de manera global [5].

3 Escenarios de trabajo

Para poder analizar el comportamiento de las redes MANET con respecto a la implementación del estándar 802.11e que provee QoS, fue necesario el uso de un simulador de redes. La herramienta de simulación usada fue Network Simulator 2 (NS-2) versión 2.28, con el parche de 802.11e [6].

Se estudiaron dos parámetros en las transmisiones para el estudio de QoS en los ambientes MANET, el *retardo extremo a extremo* y la *sobrecarga*.

Respecto a los escenarios de simulación, se trabajaron con cuatro escenarios de 20y 100 nodos, esta variación en la granularidad de los escenarios permitió el análisis del impacto de los distintos niveles de sobrecarga de paquetes de ruteo.

La cantidad de transmisiones en las simulaciones es proporcional a la cantidad de nodos. Se usó la relación de n/2 cantidad de transmisiones activas, donde n es la cantidad de nodos que se crearon. De esta forma, y siempre que n sea un número par todos los nodos estarán involucrados en al menos un flujo de datos punto a punto.

Con respecto a las transmisiones, se consideraron dos tipos. El primer tipo de transmisión es de voz 64 Kb/s, este flujo de datos está destinado a los nodos con mayor prioridad (nodos cero y n-1). Este flujo de datos se marcará con un ID particular para poder ubicarlo en los archivos resultantes de la simulación. Para el segundo tipo de transmisiones, se consideró que el resto de los nodos corriera una aplicación del tipo CBR de video (*Constant Bit Rate* – Tasa de bits constantes) de 4Mbits/s.

El protocolo de ruteo seleccionado para realizar las simulaciones es AODV (Ad-hoc *On-demand Distance Vector Routing*) debido a que el retardo que se produce por el rearmado de las tablas de ruteo tiende a estabilizarse cuando la granularidad de la red es alta (superior a los 10 y 15 nodos) [7], lo cual es propicio para los escenarios que se manejan en las simulaciones

Las simulaciones se realizaron durante 2000 segundos sobre dos modelos de movilidad; *RWPM (Random Waypoint Mobility Model)* y *RWKM (Random Walk Mobility Model)* [8].

En cuanto a las aplicaciones, la transmisión de voz, marcada con un ID especial, comienza a ejecutarse en el segundo 1.0 (el nodo 0 comienza la transmisión de datos) y se detiene 5 segundos antes de que el tiempo de simulación llegue al final. Con las demás transmisiones el mecanismo es el siguiente: se toma el número de nodo que tiene la tarea de enviar datos (dicho número está entre los números (1, n/2 - 1)) y a ese número se le suman 5.0 segundos. El resultado de esa operación indica el momento dentro de la simulación en que el nodo comenzará a transmitir. Al igual que la

transmisión marcada, las demás finalizan 5 segundos antes que el tiempo de simulación llegue a su fin.

4 Resultados Obtenidos

A continuación se muestran los resultados de las simulaciones, donde se evalúa el impacto de la implementación de 802.11e en el retardo extremo a extremo y la sobrecarga de paquetes de ruteo, también se analiza el comportamiento de esta implementación en escenarios con 20 y 100 nodos para evaluar si la granularidad posee efectos en el desempeño de la red.

4.1 Análisis del retardo

La Figura 1 muestra que el retardo es mucho menor cuando se aplica QoS a través de 802.11e. El retardo promedio sin QoS es de 10,50 segundos (±7,50 segundos), mientras que en el caso de una red con QoS se registra un promedio de 0,010 segundos (± 0,014 segundos). Los valores registrados cuando no se aplica QoS son muy dispares durante toda la simulación.

Figura 1: Retardo con 20 Nodos, con RWKM

La Figura 2 muestra los resultados cuando el tipo de movimiento es RWPM. Aquí, el retardo promedio cuando no se implementó QoS es de 2,89 segundos (±1,30 segundos), en el caso contrario es de 0,003 (±0,004 segundos). Al igual que con el modelo RWKM, y aunque el retardo promedio es mucho más bajo, la inestabilidad del tiempo de retardo está presente también cuando el tipo de movimiento es RWPM.

Figura 2: Retardo con 20 Nodos, con RWPM

Como una primera conclusión, se puede observar que el modelo de movilidad afecta el retardo extremo a extremo cuando no se aplica QoS. El tipo de movimiento RWPM produce menores retardos en transmisiones sin QoS que el modelo RWKM. Si la red tiene implementado QoS, el modelo de movilidad no produce un gran impacto. La Figura 3 muestra que el retardo es, en general menor a 1 segundo.

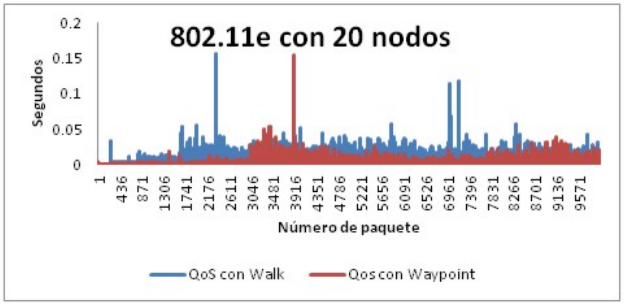


Figura 3: Comparación de QoS según la movilidad para 20 nodos

En caso de un escenario con 100 nodos, o sea, alta tasa de granularidad, se observa que el retardo se vuelve muy inestable en cualquiera de los dos tipos de movimiento cuando no se aplica 802.11e. En la Figura 4, el retardo promedio registrado fue de 11,13 segundos (±8,05 segundos) sin QoS. Mientras que cuando se aplica QoS el tiempo de retardo promedio es de 2,100 segundos (±1,266 segundos).

Figura 4: Retardo con 100 Nodos, con RWKM

La Figura 5 presenta los retardos cuando el movimiento es Random Waypoint. El tiempo promedio cuando no se aplica QoS en este caso 7,03 segundos (±6,37 segundos). De igual manera que en el caso anterior, se puede observar que la inestabilidad de los retardos, cuando no se aplica 802.11e, está muy marcada. El tiempo promedio cuando se aplica QoS es de 1,503 segundos (±1,204 segundos).

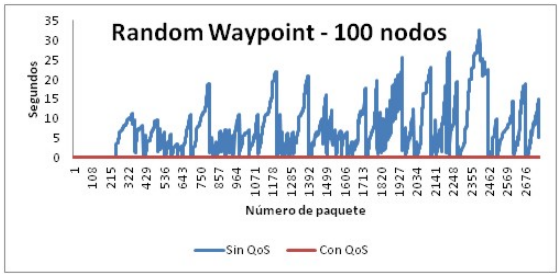


Figura 5: Retardo con 100 Nodos, con RWPM

La Figura 6 muestra la comparativa entre los tipos de movimiento cuando se aplica QoS. Se puede percibir que el tipo de movimiento RWPM genera más inestabilidad que el RWKM. Así mismo, el retardo más alto se lo registró con el movimiento de tipo Walk. En líneas generales el aumento excesivo de la granularidad aumenta el retardo en la red, aún teniendo QoS, para cualquier tipo de movimiento.

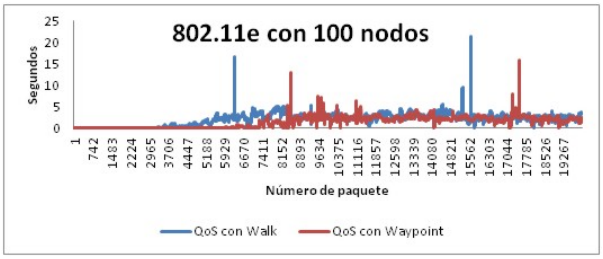


Figura 6: Comparación de QoS según la movilidad para 100 nodos

4.2 Análisis de la sobrecarga

En la Figura 7 se observa la sobrecarga de paquetes de ruteo cuando la cantidad de nodos es 20 y el tipo de movimiento es de Random Walk. Se puede ver que la sobrecarga durante toda la simulación es inestable para la red sin QoS implementado, mientras que para aquella que tiene implementada 802.11e la sobrecarga es estable y en general es baja.

Para el caso de la red sin QoS se registró una cantidad promedio de paquetes de ruteo de 66 paquetes (±43 paquetes), lo cual indica una gran inestabilidad en la sobrecarga de la red. Para su contraparte con QoS se registró una cantidad promedio de 5 paquetes (±2 paquetes), lo cual confirma lo expuesto en el primer párrafo.

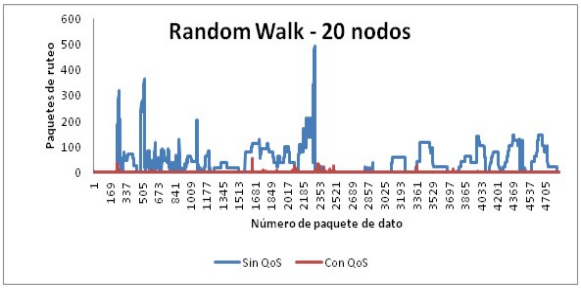


Figura 7: Sobrecarga con 20 Nodos, con RWKM

En el caso de la red cuyo movimiento es el Random Waypoint, Figura 8, se observa que la sobrecarga sigue siendo mayor en la red sin 802.11e. La sobrecarga continúa siendo inestable pues se registró una cantidad promedio de 35 paquetes (±17 paquetes). En cuanto a la red con QoS se registró una cantidad promedio de 3 paquetes (±1 paquetes).

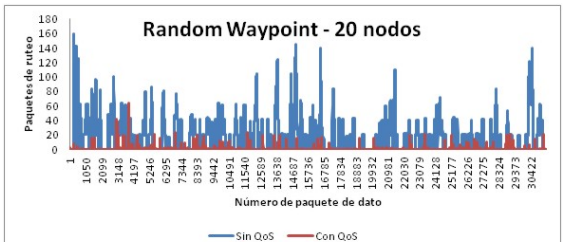


Figura 8: Sobrecarga con 20 Nodos, con RWPM

Como se muestra en la Figura 9, al comparar el impacto del tipo de movimiento en las redes con QoS, se observa que el modelo de movilidad con la granularidad estudiada, no afecta en gran medida la sobrecarga de la red cuando se implementa 802.11e.

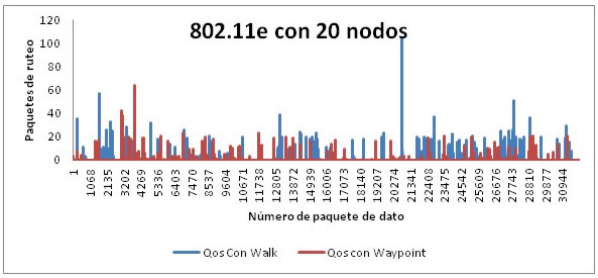


Figura 9: Comparación de QoS según la movilidad para 20 nodos

En la Figura 10, con el movimiento del tipo Random Walk, se evidencia que la sobrecarga es muy inestable cuando no se aplica QoS. El promedio de paquetes de ruteo en el caso de la red sin QoS implementado es de 5087 paquetes (±3304 paquetes), mientras que para la red con QoS el promedio de paquetes de ruteo registrado es de 2880 paquetes (±1354 paquetes).

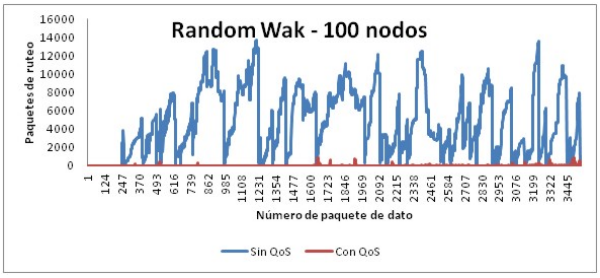


Figura 10: Sobrecarga con 100 Nodos, con RWKM

En el caso de las redes con movimiento Random Waypoint, Figura 11, el promedio de paquetes de ruteo que se registró cuando el estándar 802.11e no estaba implementado fue de 4122 paquetes (±3104 paquetes), mientras que cuando no se implementó se obtuvo un promedio de 2561 paquetes (±1294 paquetes).

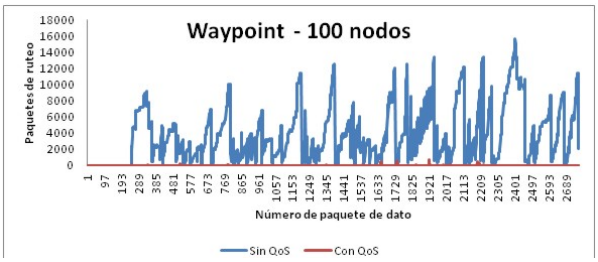


Figura 11: Sobrecarga con 100 Nodos, con RWPM

La Figura 12 muestra la diferencia entre las redes con QoS implementado según el tipo de movimiento. Si bien no se puede observar claramente cuál es más inestable, el hecho de que la desviación estándar sea mayor que el promedio en el caso de la red con movimiento Random Walk, es un indicador de que esta es la que presenta mayor inestabilidad. Aunque en líneas generales la red con mayor sobrecarga es aquella cuyo movimiento es del tipo Random Waypoint.

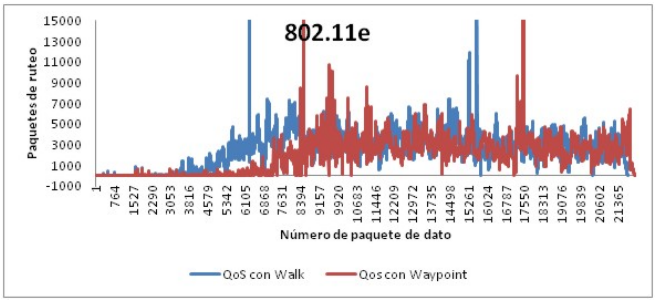


Figura 12: Comparación de QoS según la movilidad para 100 nodos

5 Conclusiones

Respecto al retardo, se puede concluir que la implementación del estándar 802.11e produce mejores resultados en las redes que lo implementan. Cualquiera sea el tipo de movilidad, y la cantidad de nodos, el retardo es siempre más bajo y más estable cuando el tráfico pertenece a una red con 802.11e. Aun cuando la granularidad es alta el retardo promedio en las redes con QoS es menor a 1 segundo. Siempre que la granularidad de la red sea menor a 100 nodos, el retardo promedio será muy bajo, no importa qué tipo de movimiento tienen los nodos.

En cuanto a la estabilidad de los tiempos de retardo se puede concluir que en los escenarios donde no se implementó el estándar 802.11e, el retardo fue muy inestable. Es decir, que sin importar el tipo de movimiento que tengan los nodos ni la cantidad, el retardo varía en forma significativa. Lo anterior permite afirmar que entornos con características de movimiento y granularidad como los que se simularon, no son aptos para aplicaciones en tiempo real sin una administración acorde de QoS a nivel MAC, las cuales son altamente sensibles a los cambios en el throughput de la red. Para las redes con 802.11e, la estabilidad cuando la granularidad es baja no es un factor

importante, debido a que, si bien en la mayoría de los casos la misma estaba presente, el retardo promedio es tan bajo (inferior al segundo) que la estabilidad o inestabilidad no afecta demasiado la eficiencia de las aplicaciones que son sensibles al mismo.

Con respecto a la sobrecarga, el estándar 802.11e marca una gran diferencia en la recarga entre las redes que lo implementan de aquellas que no lo hacen. Mientras que la granularidad sea baja la sobrecarga se mantiene estable y no alcanza grandes valores cuando se implementa el estándar de QoS a nivel de MAC. Cuando la granularidad de la red es baja, y se implementa QoS, la sobrecarga de la red es estable y se mantiene en bajos niveles. Para una taza de granularidad alta, y si las aplicaciones son del tipo CBR, la sobrecarga se vuelve inestable y alta. Esto permite concluir que en el caso de redes con alta granularidad, las aplicaciones sensibles al ancho de banda se verán seriamente afectadas, debido a que la sobrecarga no sólo es alta sino que además es inestable. Las redes con alta granularidad, sin importar el tipo de movimiento que tengan los nodos, no son aptas para aplicaciones sensibles al ancho de banda, debido a que existe una gran sobrecarga de ruteo, y esta es muy inestable. De lo anterior se puede concluir que los entornos móviles con alta granularidad no son aptos para aplicaciones como video llamadas, juegos en línea, streaming de video HD, etc., que son, como se mencionó, sensibles al ancho de banda.

En lo que respecta al efecto de los modelos de movilidad analizados, cuando no se aplica QoS a nivel de la subcapa MAC, el modelo de movilidad tiene un efecto directo y drástico en los retardos de los paquetes de datos. El modelo de movilidad junto con la granularidad fueron los causantes de la gran inestabilidad que se detectó en los retardos en las simulaciones. En general, el modelo Random Walk fue el que generó mayores y más inestables retardos que su contraparte, el Random Waypoint.

6 Bibliografía

[1] Michel Barbeau y EvangelosKranakis. “Principles of Ad-Hoc Networking”.John Wiley and Sons – 2007.
[2] Kim, Anbin. “QoS support for advanced multimedia systems”. Information Networking (ICOIN), 2012 International Conference.Page(s): 453 - 456
[3] Murazzo, Rodríguez, Vergara, Carrizo, González, Grosso. “Administración de QoS en ambientes de redes de servicios convergentes”. WICC 2013, Parana, Entre Rios, Argentina.
[4] Robert Wójcik. “Flow Oriented Approaches to QoS Assurance”. Journal ACM Computing Surveys (CSUR).Volume 44 Issue 1, January 2012 .Article No. 5.
[5] Díaz, Marrone, Barbieri, Robles. “Ruteo en redes ad-hoc”. WICC 2010, Calafate, Santa Cruz, Argentina.
[6] Wietholter, Hoene. “Design and Verification of an IEEE 802.11e EDCF Simulation Model in ns-2.26”.Techical University Berlin Telecommunication Networks Group (2003).
[7] Marrone, Robles, Murazzo, Rodríguez, Vergara. "Administración de QoS en MANET”. WICC 2011 – Rosario, Santa Fe, Argentina.
[8] Mohapatra, Panda. "Implementation and Comparison of Mobility Models In Ns-2”. National Institute of Technology, Rourkela 2009.

Caso de estudio de comunicaciones seguras sobre redes móviles ad hoc

Sergio H. Rocabado Moreno[1], Daniel Arias Figueroa[1], Ernesto Sánchez[1], Javier Díaz[2]
[1]C.I.D.I.A. – Centro de Investigación y Desarrollo en Informática Aplicada (UNSa)
[2]L.IN.T.I. – Laboratorio de Investigación en Nuevas Tecnologías Informáticas (UNLP)
[1]srocabad@cidia.unsa.edu.ar, [1]daaf@cidia.unsa.edu.ar, [1]esanchez@cidia.unsa.edu.ar, [2]jdiaz@unlp.edu.ar

Resumen. En este trabajo se presenta el estudio de un caso de integración de una MANET desplegada en zona remota a una red de infraestructura, buscando un equilibrio entre el nivel de seguridad y el consumo de recursos como la energía y el ancho de banda. Se implementaron canales de comunicación extremo a extremo, entre un nodo de la MANET y el servidor de infraestructura. Inicialmente se efectuaron pruebas inyectando tráfico de datos sobre un canal "no seguro", con la finalidad de obtener métricas de referencia como latencia, throughput y consumo de energía. Luego se configuraron canales "seguros" sobre los que se realizaron las mismas pruebas utilizando protocolos como IPSEC y SSL/TLS. Las métricas obtenidas utilizando canales seguros fueron comparadas con las de referencia para determinar las diferencias de consumo de recursos introducidas por la seguridad.

Palabras Clave: MANET, Latencia, Energía, Throughput, IPSec, SSL, TLS, Bluetooth, GPRS.

1. Introducción

Una red móvil ad-hoc o MANET (Mobile Ad hoc NETworks en inglés) [1] es una colección de nodos inalámbricos móviles que se comunican de manera espontánea y autoorganizada constituyendo una red temporal sin la ayuda de ninguna infraestructura preestablecida (como puntos de acceso WiFi o torres de estaciones base celulares) ni administración centralizada.

Por sus características las MANET constituyen una tecnología ideal para facilitar servicios de comunicación a dispositivos móviles en zonas remotas donde no es posible montar y configurar redes de infraestructura debido a inconvenientes físicos y recursos limitados, como la energía y/o cobertura de red celular.

Una ventaja adicional de este tipo de redes es la posibilidad de integrarlas a redes de infraestructura con diferentes fines, entre otros podemos mencionar, el acceso a Internet y a sistemas de información de una intranet desde los dispositivos móviles que forman parte de la MANET. Las características intrínsecas de este tipo de redes (incluyendo autoconfiguración, ausencia de infraestructura, movilidad de los nodos, topología dinámica, ancho de banda limitado, falta de seguridad, conservación de

energía, entre otras), plantean exigencias que deben resolverse antes de realizar la integración [2].

Nuestra investigación se enfoca en los siguientes aspectos:

- **Seguridad**. Las redes móviles utilizan un medio compartido (aire) para transmitir los datos y se encuentran expuestas a "ataques" o accesos no autorizados, y por esta razón se hace necesario utilizar protocolos de seguridad que permitan una integración "segura" de los dispositivos móviles a la red de infraestructura, garantizando el cumplimiento de los siguientes aspectos de seguridad: Confidencialidad, integridad, autenticación y no repudio.
- **Conservación de Energía**. Los dispositivos móviles que conforman la MANET tienen capacidad limitada de energía y pocas posibilidades para recarga de baterías cuando se encuentran en zonas remotas de recursos energéticos limitados, por lo tanto se debe optimizar el consumo de energía.
- **Ancho de banda limitado**. La integración de una MANET en zona remota a una red de infraestructura requiere el uso de la red celular. En este tipo de zonas la cobertura de red celular es muy baja y debido a ello proporciona un ancho de banda reducido y variable.

Los tres aspectos son importantes y están directamente relacionados, se debe tener en cuenta que la implementación de un protocolo de seguridad implica un consumo de energía adicional por tres motivos: 1. Se incrementa el uso de CPU y memoria para realizar cálculos, 2. Se generan encabezados adicionales (overhead) que deben ser transmitidos y 3. Se intercambian mensajes para el establecimiento de canales de comunicación seguros.

Por otra parte, la implementación de niveles de seguridad elevados implica un aumento en el consumo de energía en los nodos móviles que reduce drásticamente el tiempo de vida de la red, y un consumo adicional de ancho de banda que puede comprometer el normal funcionamiento de las aplicaciones. Debido a estas razones se hace necesario establecer un compromiso entre seguridad y consumo de recursos.

En este trabajo se presenta el estudio de un caso de integración de una MANET, desplegada en una zona remota, a una red de infraestructura. El objetivo principal es el de proporcionar, a los nodos de la red ad hoc, acceso "seguro" a un servidor de la red de infraestructura, sin comprometer recursos como ancho de banda y energía que son limitados en la zona de despliegue. Para ello, se implemento un escenario de pruebas que comprende el despliegue de una MANET en zona remota y la integración de la misma a una red de infraestructura a través de la red celular. Sobre el escenario propuesto se establecieron canales de comunicación extremo a extremo, entre un nodo de la MANET y un servidor de infraestructura. Inicialmente, se realizaron pruebas inyectando tráfico de datos sobre un canal "no seguro" para obtener valores de referencia para latencia, throughput y consumo de energía. Luego, se efectuaron las mismas pruebas utilizando canales de comunicación "seguros" configurados sobre protocolos IPSEC y SSL/TLS. Los resultados obtenidos utilizando canales "seguros" fueron comparados con los valores de referencia para determinar las diferencias de consumo de recursos. Las desviaciones que surgieron de estas comparaciones, permitieron:

- Establecer el consumo adicional de recursos generado por el uso de protocolos seguros.

- Realizar un estudio comparativo de rendimiento, entre diferentes configuraciones de protocolos de seguridad.
- Determinar que protocolo seguro se adapta mejor a este tipo de entornos.

2 Trabajos previos del grupo de investigación

En [3], se desplegó un escenario de pruebas *indoor* sin considerar condiciones externas como distancia, interferencias y otras. Se efectuaron mediciones sobre un canal "no seguro" y luego sobre un canal "seguro", el aseguramiento del canal se implemento utilizando diferentes configuraciones del protocolo IPSec en modo transporte (extremo a extremo). En los resultados se presentan gráficos comparativos de consumo de energía entre las diferentes configuraciones de seguridad.

En [4], continuamos con esta línea de investigación, utilizando IPSec para el aseguramiento del canal de comunicaciones, esta vez sobre un escenario de pruebas *outdoor* afectado por factores externos que disminuyen el rendimiento e incrementan el consumo de recursos en los nodos de la red ad hoc. En el desarrollo de la publicación, se fundamenta la elección de Bluetooth como tecnología de soporte para la formación de la MANET remota y de GSM/GPRS para la integración de la misma a la red de infraestructura. Entre los resultados se presentan gráficos que muestran el consumo de energía para cada configuración de canal y la distribución del consumo entre los siguientes ítems: Establecimiento de sesión, encriptación, autenticación y transmisión.

En [5], se describe una experiencia del uso de MANETs en zonas rurales de recursos limitados (energía y ancho de banda). En el trabajo de campo realizado, se desplegaron MANETs de bajo consumo en escuelas rurales, con la finalidad de facilitar a docentes y alumnos el acceso a contenidos m-learning instalados en un servidor de infraestructura. Se consiguió mantener el rendimiento de la MANET dentro niveles aceptables de eficiencia y sin comprometer los recursos, lo que posibilito un funcionamiento correcto de las estrategias de m-learning en este tipo de zonas. Entre las conclusiones de esta publicación se destaca la siguiente: "El uso de las MANETs es efectivo y eficiente para el desarrollo de experiencias de m-learning en zonas de recursos energéticos limitados".

3. Escenario de pruebas

En la Figura 1 se observa la representación gráfica del escenario implementado para realizar las pruebas y mediciones. En el mismo se conecta una MANET, desplegada en zona remota, a la Intranet del campus universitario de la UNSa, a través de la red celular (GSM/GPRS). Los dispositivos móviles (nodos) de la MANET se conectan al servidor "testing" utilizando un canal de comunicación TCP/IP extremo a extremo (end to end). El tráfico entre el nodo móvil y el servidor se gestiona a través de uno de los nodos que actúa como Gateway entre la MANET y la red celular. Este nodo es el encargado de enviar los paquetes de datos hacia los routers de la red celular; desde

donde y a través de Internet son encaminados a la intranet para ser entregados al servidor.

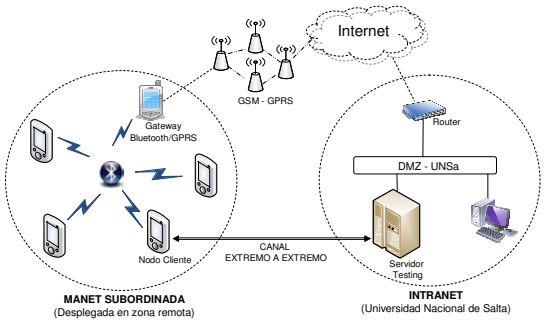


Figura 1. Escenario de pruebas

La conexión del dispositivo móvil cliente al punto de acceso a la red (NAP - Network Access Point) se realizó utilizando el perfil PAN (Personal Área Network) [6] del estándar Bluetooth [7]. El punto de acceso a la red se configuro sobre el nodo Gateway utilizando la funcionalidad "Bluetooth Tethering" de Android, que utiliza el Framework netfilter e iptables para implementar un puente entre la PAN bluetooth y la red GSM/GPRS [8].

El canal de comunicación provee comunicación TCP/IP, extremo a extremo, entre el nodo cliente y el servidor "Testing". En el trayecto los datagramas IP son encapsulados en BNEP [9] por la red Bluetooth y GTP [10] por la red GPRS.

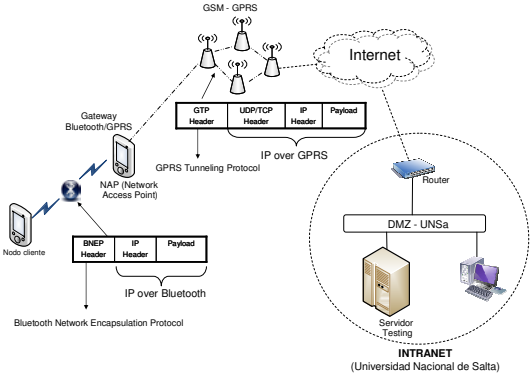


Figura 2. Comunicación extremo a extremo

La Figura 2 ilustra el envío de un datagrama IP desde el nodo móvil hasta el servidor de la intranet siguiendo los siguientes pasos:

1. El nodo móvil envía el datagrama IP, encapsulado en BNEP [9], al punto de acceso a la red (NAP).
2. El NAP transmite el datagrama al SGSN de la red GPRS, desde donde viaja al GGSN encapsulado en GTP [10].
3. El GGSN re-envía el datagrama a Internet, por donde viaja hasta llegar al router frontera de la red destino.

4. El router frontera de la red destino encamina el datagrama hacia el servidor, encapsulado en una trama Ethernet.

3.1 Configuración del nodo cliente

La configuración del dispositivo móvil con el cual se realizaron las pruebas, es la siguiente:

Equipo:	Samsung I9300 Galaxy S III
CPU:	Quad-core 1.4 GHz Cortex-A9
Chipset:	Exynos 4412 Quad
RAM:	1024MB RAM.
SO:	Android OS ver. 4.1.2 (Jelly Bean)
Root:	SI
Bluetooth:	ver 3.0
Bateria:	Lítio-íon, 2100 mAh, 3.7 v.

Este equipo fue especialmente preparado para minimizar el consumo de batería, se procedió entonces a: desinstalar las aplicaciones no indispensables para su funcionamiento, deshabilitar dispositivos de hardware no utilizados en las pruebas y habilitar el modo de bajo consumo.

4. Métricas y aplicaciones elegidas para efectuar las mediciones

Para medir el rendimiento del canal de comunicaciones extremo a extremo, entre el nodo cliente y el servidor, se eligieron las siguientes métricas: Latencia, Throughput y Consumo de energía. En la tabla 1 se muestran las aplicaciones, del lado del cliente y del lado del servidor, que fueron utilizadas para obtener las métricas.

Tabla 1. Aplicaciones utilizadas para obtener las métricas

Métrica	Aplicación Cliente	Aplicación Servidor
Latencia ICMP	Busybox Ping [11]	Windows Stack TCP/IP
Latencia HTTP	HTTPing [12]	HTTP Server (IIS6)
Throughput TCP	Iperf Client [13]	Iperf Server
Throughput HTTP	Busybox Wget [11]	HTTP Server (IIS6)
Throughput FTP	Busybox Wget [11]	FTP Server (Filezilla)
Consumo de energía	Powertutor [14]	----------

5. Configuraciones de canal

Los canales fueron divididos en 2 grupos: 1. Canal no seguro y 2. Canal seguro basado en VPN (L2TP/IPSEC, OPENVPN SSL/TLS y OPENVPN SSL/TLS con compresión LZO). La tabla 2 resume las aplicaciones utilizadas para la implementación de los canales.

Para establecer un canal de comunicación NO seguro, alcanza con brindar transporte IP entre el nodo cliente y el servidor (ver figura 2), mientras que para el establecimiento de canales seguros se requiere el uso de protocolos seguros como IPSec y SSL/TLS.

Tabla 2. Configuraciones de seguridad (Cliente/Servidor)

CANAL	Protocolo seguro	Cliente (Android)	Servidor (Windows)
No seguro	Ninguno	N/A	N/A
Seguro	IPSEC	CLIENTE L2TP/IPSEC	RRAS L2TP/IPSEC
	SSL/TLS	OPENVPN Client [15]	OPENVPN Server [16]
	SSL/TLS/LZO	OPENVPN Client [15]	OPENVPN Server[16]

El canal seguro IPSec se implemento utilizando el protocolo L2TP encapsulado en IPSec [17]. IPSec se configuro en modo transporte utilizando una asociación de seguridad (SA) entre el nodo cliente y el servidor. Elegimos: 1. El algoritmo RSA para la autenticación mutua, entre el nodo cliente y el servidor 2. El algoritmo SHA1 para calcular el código de autenticación de mensaje (MAC), utilizado para verificar la integridad de los mensajes y 3. El algoritmo de encriptación 3DES para la confidencialidad.

El canal seguro SSL/TLS [18] se configuro utilizando la aplicación OPENVPN con y sin compresión LZO [19]. Elegimos: Autenticación RSA de tipo desafío respuesta para el servidor y autenticación RSA para el nodo cliente, HMAC-SHA1 para la integridad y 3DES para la confidencialidad.

La gestión de lo certificados utilizados por los protocolos seguros, fue realizada por una CA montada en el servidor "Testing".

Tabla 3. Protocolos y algoritmos utilizados para la implementación de canales seguros

Canal	Protocolo	Autenticación	Cifrado	Integridad	Compresión
NO Seguro	IP	n/a	n/a	n/a	n/a
Seguro L2TP/IPSEC	IPSEC	RSA (mutual)	3DES (168 bits)	HMAC-SHA1 (160 bits)	n/a
Seguro OPENVPN	SSL/TLS	RSA (Servidor, Cliente)	3DES (168bits)	HMAC- SHA1 (160 bits)	n/a
Seguro OPENVPN	SSL/TLS	RSA (Servidor, Cliente)	3DES (168 bits)	HMAC- SHA1 (160 bits)	LZO

6. Mediciones realizadas

En la tabla 4 se presentan las mediciones efectuadas para cada configuración de canal y el mecanismo utilizado para generar trafico entre el cliente y el servidor.

Tabla 4. Mediciones y mecanismos utilizados

Medición	Mecanismo utilizado para generar tráfico
Latencia ICMP	Echo Request/Reply (32 bytes)
Latencia HTTP	HTTP GET
Throughput TCP y	Inyección de tráfico TCP aleatorio (1024 Kbytes)

consumo de energía	
Throughput HTTP y consumo de energía	Descarga de archivo (de 1024 Kbytes) utilizando HTTP.
Throughput FTP y consumo de energía	Descarga de archivo (de 1024 Kbytes) utilizando FTP.

Las mediciones se realizaron de manera automática, utilizando aplicaciones (figura 3) que se ejecutaron de manera continua durante 7 días en la franja horaria 6:00 am a 11:00 pm, de esta manera fueron contemplados diferentes niveles de carga de la red GPRS. Los resultados obtenidos se promediaron para determinar el valor final de cada medición.

MEDICIONES	CONFIGURACION DE CANAL			
	NO SEGURO	L2TP IPSEC	OPENVPN (SSL/TLS)	OPENVPN LZO (SSL/TLS)
Latencia ICMP (Busybox Ping)	SI	SI	SI	SI
Latencia HTTP (HTTPing)	SI	SI	SI	SI
Throughput TCP (Iperf)	SI	SI	SI	SI
Throughput HTTP (Busybox Wget)	SI	SI	SI	SI
Throughput FTP (Wget - ANDftp)	SI	SI	SI	SI

Consumo de energía (PowerTutor)

Figura 3. Resumen de configuraciones de canal implementadas, mediciones realizadas y aplicaciones utilizadas para medir.

El consumo de energía en el nodo cliente se midió utilizando la aplicación PowerTutor [14], esta herramienta permite estimar el consumo de energía en tiempo real y por proceso utilizando el modelo de consumo de energía descrito en [20] .

6.1 Metodología de medición

A continuación se enumeran los pasos necesarios para efectuar una medición:

- Establecer comunicación extremo a extremo, según la configuración de canal utilizada (ver tabla 3).
- Ejecutar la aplicación Powertutor.
- Arrancar el monitoreo de consumo de energía (“Start Profiler”)
- Generar tráfico entre el Cliente y el Servidor, el mecanismo depende de la medición realizada (ver tabla 4)
- Detener Powertutor (“Stop Profiler”)
- Guardar el “log” de Powertutor (Pulsar Menú -> Save Log).
- Copiar el “log” generado por “Powertutor”.
- Copiar el “log” generado por la aplicación utilizada para la medición.
- Analizar y procesar los archivos de logs.
- Promediar resultados.

7. Resultados

A continuación se presentan algunos de los resultados obtenidos, utilizando gráficos que resumen los aspectos estudiados.

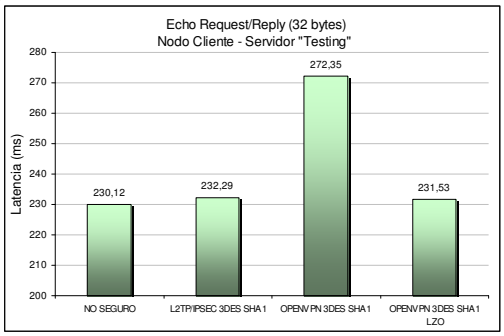


Figura 4. Latencia ICMP echo request/Reply

En la figura 5 se presenta el Throughput alcanzado una descarga de archivo de 1024 Kbytes, utilizando los protocolos HTTP y FTP. Se observa que HTTP alcanza un rendimiento ligeramente superior a FTP en todos los casos.

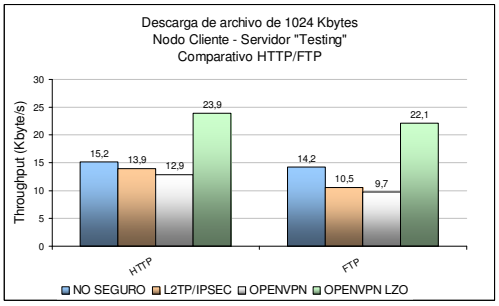


Figura 5. Throughput de descarga HTTP y FTP

El grafico de la figura 6 muestra el Throughput, obtenido por la aplicación Iperf para cada configuración de canal, y la cantidad de energía (en joules) consumida por iperf para efectuar la prueba. Comparando los resultados obtenidos para el canal no seguro y el canal seguro con compresión, se observa que la compresión mejora considerablemente el Throughput (~ 400%) e introduce un consumo adicional de energía (~ 200%).

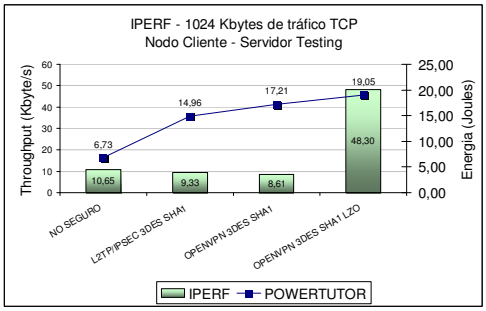


Figura 6. Rendimiento TCP/Iperf (throughput vs energía consumida)

8. Conclusiones y trabajos futuros

La seguridad implica un consumo adicional de recursos que puede variar dependiendo de los protocolos que se elijan para el establecimiento de un canal seguro, en los gráficos comparativos presentados se visualiza que la opción de seguridad basada en SSL/TLS con compresión (3DES - SHA-1 - LZO) es la que mayor energía consume, triplicando el consumo de un canal no seguro.

Se evidencia que el incremento en el consumo de energía introducido por la compresión es proporcionalmente bajo en relación a la mejora de throughput alcanzada. Esto hace, que la compresión sea una opción a considerar en escenarios donde se requiera mejorar el rendimiento del ancho de banda y la energía no sea un factor critico.

Observamos que el protocolo IPSec consigue un mejor aprovechamiento del ancho de banda y menor consumo de energía, comparado con el protocolo SSL/TLS sin compresión.

El uso de canal seguro en lugar de un canal no seguro, introduce una disminución en el rendimiento del ancho de banda (entre un 10% y 20%) y un incremento en el consumo de energía (entre un 100% y 200%). La elección de un nivel de seguridad en los nodos ad hoc dependerá de las posibilidades de recarga de energía y del ancho de banda disponible en la zona de despliegue de la MANET.

Para continuar con esta línea de investigación tenemos previsto:

- Utilizar dispositivos con interfaces Bluetooth 4.0 de bajo consumo.
- Efectuar pruebas variando la potencia de transmisión del nodo cliente y la distancia entre el nodo cliente y el nodo Gateway.
- Incorporar compresión DEFLATE al protocolo IPSEC y comparar los resultados con los obtenidos para la compresión LZO.
- Utilizar herramientas de simulación para modelar el comportamiento aleatorio y la congestión de la red GPRS.
- Estudiar la distribución del consumo de energía entre los componentes de un protocolo seguro: Intercambio de claves, autenticación, integridad, encriptación y transmisión de datos.

Referencias

1 IETF. *MANET Active Work Group*. http://tools.ietf.org/wg/manet

2 CORDEIRO DE MORAIS, Carlos and AGRAWALL Dharma. (2011). Integrating MANETs, WLANs and Cellular Networks. In World Scientific Publishing (Ed.), *Ad Hoc and Sensor Networks - Theory and Applications* (pp. 587-620). Singapore: World Scientific Publishing.

3 ROCABADO, Sergio; SANCHEZ, Ernesto; DIAZ, Javier y ARIAS FIGUEROA, Daniel. (2011). *Integración Segura de MANETs con Limitaciones de Energía a Redes de Infraestructura*. Paper presented at the CACIC 2011, La Plata - Buenos Aires - Argentina. http://sedici.unlp.edu.ar/handle/10915/18771

4 ROCABADO, Sergio; SANCHEZ, Ernesto; DIAZ, Javier y ARIAS FIGUEROA, Daniel. (2012). *Integración Segura de MANETs, desplegadas en zonas de recursos*

limitados, a Redes de Infraestructura. Paper presented at the CACIC 2012, Bahia Blanca - Buenos Aires - Argentina. http://sedici.unlp.edu.ar/handle/10915/23762

5 ROCABADO, Sergio; HERRERA, Susana y Otros. (2013). *M-LEARNING EN ZONAS DE RECURSOS LIMITADOS*. Paper presented at the TE&ET 2013, Santiago del Estero - Argentina.

6 CORDEIRO DE MORAIS, Carlos and AGRAWALL Dharma. (2011). Wireless PANs. In World Scientific Publishing (Ed.), *Ad Hoc and Sensor Networks - Theory and Applications* (pp. 196-258). Singapore: World Scientific Publishing.

7 SPECIAL INTEREST GROUP (SIG) Bluetooth. (2001). Specification of the Bluetooth System, tomo 2. *Bluetooth Profiles Specification Version 1.1*.

8 ETSI EN 301 344. (2000). *Digital cellular telecommunications system, General Packet Radio Service (GPRS), Service description*. Retrieved from http://www.etsi.org/index.php/technologies-clusters/technologies/mobile/gprs

9 SPECIAL INTEREST GROUP (SIG) Bluetooth. (2001). Bluetooth Network Encapsulation Protocol (BNEP) Especification.

10 3GPP. (2011). Specification 29060 - GPRS Tunneling Protocol, release 11.0. from http://www.3gpp.org/ftp/Specs/html-info/SpecVsWi--29060.htm

11 STERICSON, Stephen. (2011). BusyBox. from https://play.google.com/store/apps/details?id=stericson.busybox

12 FOLKERT van Heusden. HTTPing for Google Android mobile phones. Retrieved from http://www.vanheusden.com/Android/HTTPing

13 MAGICANDROIDAPPS.COM. (2011). Iperf for Android. from https://play.google.com/store/apps/details?id=com.magicandroidapps.iperf

14 GORDON, Mark; ZHANG, Lide and TIWANA, Birjodh. PowerTutor. University of Michigan. Retrieved from http://ziyang.eecs.umich.edu/projects/powertutor

15 SCHÄUFFELHUT, Friedrich. OpenVPN Installer. Retrieved from http://code.google.com/p/android-openvpn-installer

16 OpenVPN for Windows. (2010). from http://openvpn.net/index.php/download.html

17 PATEL, B.; ABOBA, B y Otros (2001). *L2TP/IPsec, RFC 3193*. IETF. Retrieved from http://tools.ietf.org/html/rfc3193

18 DIERKS, T.; RESCORLA, E. (2008). *The Transport Layer Security (TLS) Protocol (ver 1.2)*. IETF. Retrieved from http://tools.ietf.org/html/rfc5246

19 OBERHUMER, Markus F.X.J. (2010). LZO compression. from http://www.oberhumer.com/opensource/lzo/

20 ZHANG, Lide; TIWANA, Birjodh; QIAN, Zhiyun and WANG, Zhaoguang. (2010). *Accurate online power estimation and automatic battery behavior based power model generation for smartphones*. Paper presented at the 2010 IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Scottsdale, AZ. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5751489

Estudio del desempeño de OLSR en una red mallada inhalámbrica en un escenario real

Eduardo Rodríguez, Claudia Deco, Luciana Burzacca, Mauro Petinari
Departamento de Investigación Institucional, Facultad de Química e Ingeniería,
Universidad Católica Argentina,
2000 Rosario, Argentina
{ejrodriguez, cdeco, lucianaburzacca, mauro_pettinari}@uca.edu.ar

Abstract. El objetivo de este trabajo es analizar el comportamiento del protocolo OLSR sobre una red mallada configurada con firmware OpenWrt utilizando distintos equipos de hardware. Se presentan los resultados empíricos de varias pruebas utilizando el mismo escenario. El escenario que se presenta es una red de mundo real (no de laboratorio) con pruebas reales y no simulaciones. OpenWrt es un software perfectamente válido que puede ser utilizado en una gran variedad de dispositivos y su configuración para utilizarlo con protocolo OLSR es sencilla de realizar y no presenta problemas de funcionamiento con dicho protocolo.

Keywords: Redes Malladas Inalámbricas, Redes Mesh, Protocolos, OLSR, OpenWrt.

1 Introducción

El objetivo de este trabajo es analizar el comportamiento del protocolo OLSR sobre una red mallada configurada con firmware OpenWrt utilizando distintos equipos de hardware.

Las redes malladas inalámbricas (Wireless Mesh Networks) han tenido un gran éxito en la historia de las ciencias de la computación y de la ingeniería. Sus aplicaciones son numerosas en el dominio industrial, militar y comercial. Son en particular un dominio rápidamente creciente y esto trae muchos desafíos. En particular, un desafío difícil e inmediato es el enrutamiento efectivo debido a la volatilidad típica de tráfico en topologías complejas. Muchos estudios han intentado resolver el problema de enrutamiento mediante métodos heurísticos, pero este enfoque no proporciona los límites de cuán bien se asignan los recursos. Sin embargo, este tipo de investigación generalmente asume que el tráfico de demandas de la red es estático y conocido de antemano. Como resultado, estos algoritmos tienden a sufrir un desempeño pobre. De hecho, trabajos recientes han demostrado que el tráfico inalámbrico es muy variable y difícil de caracterizar. Comprender el impacto de la incertidumbre de la demanda en el ruteo y el diseño de algoritmos de enrutamiento para proporcionar robustez, es relativamente un problema de investigación aún incipiente.

Las redes Mesh abiertas son redes ad-hoc descentralizadas que no se basan en infraestructuras previas, como routers o puntos de acceso. En su lugar, cada nodo participa en el enrutado, siendo él mismo un router y enviando datos de otros, y de ese modo la determinación de las rutas se hace dinámicamente, basándose en la conectividad que va surgiendo. Para ello, necesitan de protocolos que viabilicen ese comportamiento.

Es de suma importancia el análisis de la perfomance de diferentes protocolos de comunicación que deben interactuar con diversos dispositivos que hacen al enlace de los

nodos de la red a los fines de establecer la integración tecnológica disponible. No menos importante es la determinación de la relación costo/beneficio de una determinada implementación. El conocimiento en tiempo real de la configuración topológica de la red, mediante el uso de distintas herramientas de hardware y software, nos permite el monitoreo del comportamiento y sus alcances. Todo ello posibilita optimizar la red para que brinde un mejor servicio. En general, la optimización se basa en lograr el mejor camino para enrutar los paquetes de datos, sin demoras o con una demora mínima en función de lograr un mejor aprovechamiento de los recursos utilizados.

En la Sección 2 se presentan algunos conceptos básicos sobre redes malladas y protocolos de ruteo. En la Sección 3 se describe el escenario, hardware y software utilizados. En la Sección 4 las pruebas realizadas. Finalmente, se presentan las conclusiones.

2 Conceptos Básicos

Una Red Mallada Inalámbrica (Mesh) es una red compuesta por nodos organizados en una topología de malla. Son redes en las cuales la información es pasada entre nodos en una forma de todos contra todos y en una jerarquía plana, en contraste a las redes centralizadas. Toda variación no prevista en el diseño, puede cambiar su topología, afectar a la distribución de carga de la red y al rendimiento general [1].

Las ventajas que presenta frente a otras redes son el bajo costo al utilizar enlaces inalámbricos, la facilidad de aumentar el área de cobertura incluyendo nuevos nodos, ya no es necesario cambiar infraestructuras como en el caso de las redes cableadas, la robustez que presenta ante fallos al disponer de rutas alternativas y la capacidad de transmisión que permiten aplicaciones a los usuarios en tiempo real de voz, video y datos. Por tanto se puede incluir un nuevo nodo en cualquier momento y lugar. Como consecuencia el costo de este tipo de redes inalámbricas es mucho menor que en las redes cableadas, ya que no hay que invertir en materiales de cableado y en estudios enfocados a la unión más óptima de los nodos. En la realidad, la topografía raramente viene en forma de anillo, línea recta o estrella. En terrenos difíciles, sean remotos, rural o urbano, donde no todos los usuarios ven uno o algunos puntos centrales, lo más posible es que el usuario solo vea a uno o más usuarios vecinos.

En una red mallada un conjunto de nodos se comunican entre sí de manera directa transmitiendo la información de nodo a nodo hasta que llega a su destino final. La información atraviesa múltiples saltos y no hay necesidad de una unidad centralizada que controle el modo de transmisión. La comunicación se realiza entre los nodos directamente. Cada nodo puede ser origen y destino de los datos o encaminar la información de otros nodos. Las redes malladas inalámbricas son robustas al tener varios caminos disponibles entre el nodo origen y el destino, de modo que el servicio no se ve afectado por la caída de un nodo o por la ruptura de un enlace.

Dado que la forma de operar que tienen estas redes consiste en que los datos pasan de un nodo a otro hasta que llegan a su destino, los algoritmos de ruteo dinámico necesitan que cada nodo comunique información de ruteo a otros nodos en la red. Cada nodo determina qué hacer con los datos que recibe, ya sea pasarlos al próximo nodo o quedárselos, dependiendo del protocolo utilizado. El algoritmo de ruteo usado siempre debería asegurar que la información tome el camino más apropiado de acuerdo a una métrica. Una métrica es

el valor por el cual los protocolos determinan cuál ruta tomar o con cuál nodo comunicarse.

Una de las debilidades y limitaciones de las redes Mesh es la latencia (el retardo de propagación de los paquetes), que crece con el número de saltos. Los efectos del retardo son dependientes de la aplicación. Por ejemplo los correos electrónicos no son afectados por grandes latencias, mientras que los servicios de voz son muy sensibles a los retardos. Otra debilidad es la disminución del rendimiento en todas las redes multisalto, esto es, a mayor número de saltos, se tiene menor rendimiento.

Con respecto al hardware, prácticamente cualquier nodo inalámbrico puede convertirse en un nodo Mesh simplemente mediante modificaciones de software.

Protocolos de Encaminamiento

La principal función de los protocolos de encaminamiento es seleccionar el camino entre el nodo fuente y destino de una manera rápida y fiable. Las redes malladas inalámbricas pueden utilizar los protocolos de encaminamiento de otras redes ya existentes, pero modificándolos para que funcionen correctamente con ellas. Si se elige esta opción, el protocolo de encaminamiento modificado debe asegurar las principales características que son el número de saltos, el rendimiento, la tolerancia a fallos, el equilibrado de carga, la escalabilidad y el soporte adaptativo.

Otra opción es diseñar un nuevo protocolo de encaminamiento para las redes malladas inalámbricas. Esta solución es más costosa ya que cuando se desarrolla un nuevo protocolo hay que probarlo, modificarlo y solucionar los fallos. Por tanto el tiempo de realización es mayor que si nos centramos en un protocolo ya experimentado.

En este trabajo utilizamos el protocolo OLSR para el encaminamiento en la red mesh dado que es uno de los más difundidos en este tipo de redes inalámbricas, a continuación una breve reseña.

OLSR: Optimized Link State Routing Protocol ([2], [3]) es un protocolo proactivo que se basa en el estado de los enlaces. Se utiliza la técnica MPR (Multipoint Relaying) que consiste en elegir un conjunto de nodos vecinos que cubran el acceso de nodos distantes a 2 saltos o más. Se adapta bien en redes con un gran número de nodos y de alta movilidad. El formato del paquete es igual para todos los datos del protocolo, así es fácil la extensión del mismo. Para saber el estado de un enlace se envían mensajes de HELLO. Cada nodo tiene asociado a cada vecino el estado del enlace. Cuando un nodo detecte la aparición de un nuevo vecino se debe incluir una nueva entrada a la tabla de encaminamiento e incluir el estado del enlace. Además si se detecta una variación en el estado de un enlace, se debe comprobar en la tabla de encaminamiento que el cambio ha sido reflejado. Si no se recibe información de un enlace durante un tiempo determinado se elimina de la tabla de encaminamiento el enlace y el vecino correspondiente. Para calcular las rutas, cada nodo contiene una tabla de encaminamiento con el estado del enlace y el nodo. El estado del enlace se mantiene gracias al intercambio de mensajes periódicos. La tabla de encaminamiento se actualiza si se detecta algún cambio en el campo de enlace, de vecino, de vecino de dos saltos o en la topología.

3 Escenario y Tecnologías Utilizadas

Se montó una red experimental distribuida en tres edificios del campus de la Universidad a los efectos de tener un campo de pruebas más parecido a la realidad de las redes mesh. En la Figura 1 se muestra la distribución del equipamiento y métricas del protocolo OLSR.

Al momento de montar la red mesh, se realizó un análisis del campo electromagnético en la frecuencia 2.4 ghz. Para esto se utilizó un analizador de frecuencia de Ubiquiti AirView2 ext. Se detectó que el canal 11 no estaba siendo utilizado por la red inalámbrica de infraestructura. A raíz de esto se eligió esta frecuencia para la red mesh.

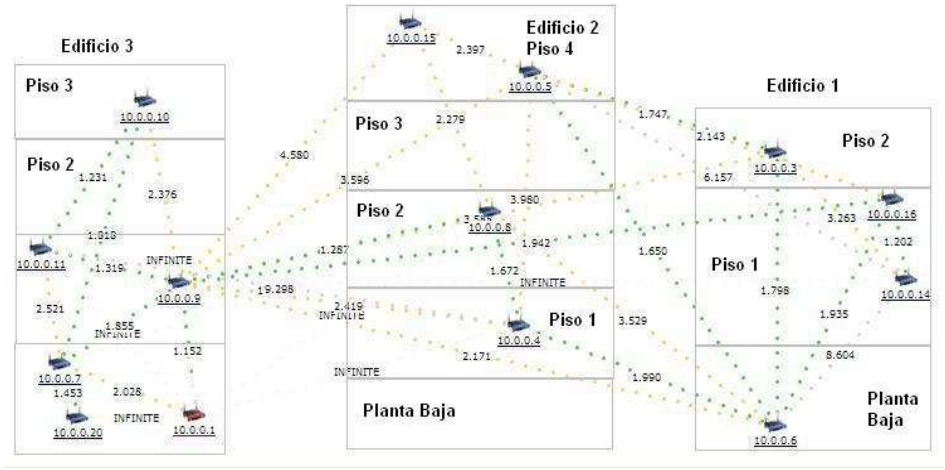


Fig. 1. Escenario

Tabla 1. Hardware utilizado.

Nro	Nombre	Hardware	SO
1	nodo01	Linksys wrt54gl	Freifunk
2	nodo02	Linksys wrt54gl	Freifunk
3	nodo03	Linksys wrt54gl	Freifunk
4	nodo04	Linksys wrt54gl	Freifunk
5	nodo05	Ubiquiti loco M2	Commotion
6	nodo06	Ubiquiti loco M2	Commotion
7	nodo07	Ubiquiti loco M2	Commotion
8	nodo08	Ubiquiti Nanostation2	Openwrt BackFire 10.03
9	nodo09	Ubiquiti Nanostation2	Openwrt BackFire 10.03
10	nodo10	Ubiquiti loco M2	Commotion
11	nodo11	Linksys wrt54gl	Freifunk
12	nodo14	TP-LINK TL743ND	Openwrt BackFire 10.03
13	nodo15	TP-LINK TL743ND	Openwrt BackFire 10.03
14	nodo16	TP-LINK TL841ND	Openwrt BackFire 10.03
15	nodo20	Linksys wrt54gl	Openwrt BackFire 10.03

En el montaje de esta red se utilizaron equipos de las marcas Linksys (WRT54GL), Ubiquiti (Nonostation 2, Nanostation Loco M2), TP-Link (TL-WR743ND, TLWR842ND) como se muestra en la Tabla 1. Se eligieron por la gran popularidad y su bajo precio.

Se utilizó como sistema de operativo OpenWrt [4] que es una distribución de Linux usada para dispositivos embebidos tales como routers personales. El soporte fue limitado originalmente al modelo Linksys WRT54G, pero desde su rápida expansión se ha incluido soporte para otros fabricantes y dispositivos. OpenWrt utiliza principalmente una interfaz de línea de comando, pero también dispone de una interfaz web en constante mejora. El soporte técnico es provisto como en la mayoría de los proyectos de Software Libre, a través de foros y su canal IRC. El desarrollo de OpenWrt fue impulsado inicialmente gracias a la licencia GPL, que obligaba a todos aquellos fabricantes que modificaban y mejoraban el código, a liberar éste y contribuir cada vez más al proyecto en general.

Como se puede ver en la tabla de dispositivos se utilizaron distintas versiones de SO:

- OpenWrt en su versión 10.03.1, que es la estable más reciente, solamente con el agregado del protocolo OLSR versión 0.6.1-3 que es la que viene standart con esa versión de openwrt

- Freifunk [5] que es una adaptación basada en OpenWrt hecha por grupos de usuarios alemanes que la utilizan para el montado de redes mesh en varias ciudades de ese país. Este sistema operativo presenta varias adaptaciones específicas para redes mesh y entre ellas una muy útil como es la graficación de los enlaces de toda la red y los valores de las métricas de OLSR para cada una como se puede ver en la Figura 1, viene instalada por defecto. La versión de OLSR es la 0.6.0.

Commotion [6] es otra adaptación de OpenWrt hecha especialmente para montado plug and play de redes mesh. Está basada en las últimas versiones de OpenWrt (10.03 en adelante) y para ser usada principalmente en equipos Ubiquiti de la serie M. Al igual que Freifunk (de hecho muchas aplicaciones vienen de esta distribución) presenta varias herramientas y utilidades para el análisis y la visualización del comportamiento de la red. También utiliza protocolo OLSR por defecto en su versión 0.6.5.4. Si uno no utiliza las configuraciones por defecto para el armado de la mesh presenta algún grado de dificultad para realizar la configuración que uno desee.

En todas estas versiones de OpenWrt se utilizaron las versiones de OLSR que se instalan por defecto desde los repositorios.

Todas estas versiones de OpenWrt utilizan un interface web que permite la configuración de todas las opciones para que la mesh funcione.

4 Pruebas realizadas

Utilizando el escenario, se realizaron pruebas para medir la efectividad del protocolo. En la ejecución de estas pruebas se utilizó el camino formado por los nodos 20, 7, 9, 16 y 14. Las métricas de rendimiento son: El tiempo de ida y vuelta (Round-trip time - RTT), Jitter, la probabilidad de error y testeo de ancho de banda.

RTT: es el tiempo que le lleva a un paquete alcanzar un nodo remoto y regresar. Está relacionado con la latencia de la conexión. Cuanto más bajo es el RTT, mejor es la conexión.

Jitter: es la variación en la latencia de paquetes recibidos de un nodo remoto. Cuanto más bajo es, mejor conexión. Es importante cuando se utiliza aplicaciones de voz sobre IP.

Probabilidad de error: Los errores en una red causan que los paquetes se pierdan, corrompan, se dupliquen o queden fuera de servicio. Cuando ocurre un error es importante saber la probabilidad con la que suceden y el tiempo entre ellos. Lo ideal es no tener errores, pero una tasa baja es aceptable.

Ancho de banda: es la tasa de transmisión de un enlace o sistema de transporte de datos y se puede definir como la capacidad de un enlace o sistema para transmitir datos. Se expresa en bit por segundo.

Nuestra herramienta principal de testeo fue iperf para todas las métricas a excepción de RTT, que se midió con ping. Iperf es una herramienta que se utiliza para hacer pruebas en redes informáticas. El funcionamiento habitual es crear flujos de datos TCP y UDP y medir el rendimiento de la red. Iperf permite al usuario ajustar varios parámetros que pueden ser usados para hacer pruebas en una red o para optimizar y ajustar la red. Puede funcionar como cliente o como servidor y puede medir el rendimiento entre los dos extremos de la comunicación, unidireccional o bidireccionalmente. Es software de código abierto y puede ejecutarse en varias plataformas incluyendo Linux, Unix y Windows. Cuando se utiliza el protocolo UDP, Iperf permite al usuario especificar el tamaño de los datagramas y proporciona resultados del rendimiento y de los paquetes perdidos. Cuando se utiliza TCP, Iperf mide el rendimiento de la carga útil. Típicamente la salida de Iperf contiene un informe con marcas de tiempo con la cantidad de datos transmitidos y el rendimiento medido.

Para medir el valor de RTT se utilizó la herramienta ping. Ping es el acrónimo de Packet Internet Groper, que significa "Buscador o rastreador de paquetes en redes". Es un utilitario que analiza el estado de la comunicación entre un host local y uno o varios remotos por medio del envío de paquetes. Se utiliza para diagnosticar el estado, velocidad y calidad de una red determinada. En nuestras pruebas siempre se utilizó como tamaño de paquete 1016.

La Figura 2 muestra los resultados de RTT obtenidos utilizando el protocolo OLSR. Los resultados muestran un patrón donde RTT se incrementa con el número de saltos.

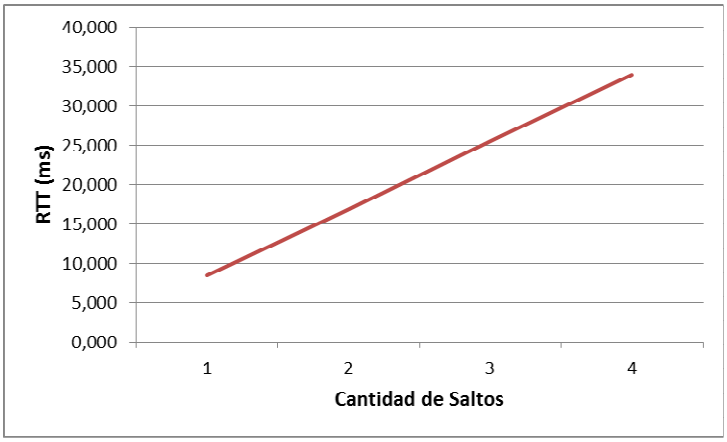


Fig. 2. Evaluación de Round-trip Time en OLSR

La Figura 3 muestra los resultados de la variación del retardo (jitter) obtenidos. A medida que aumenta el número de saltos aumenta el valor de retardo y el aumento se torna significativo para 3 y 4 saltos.

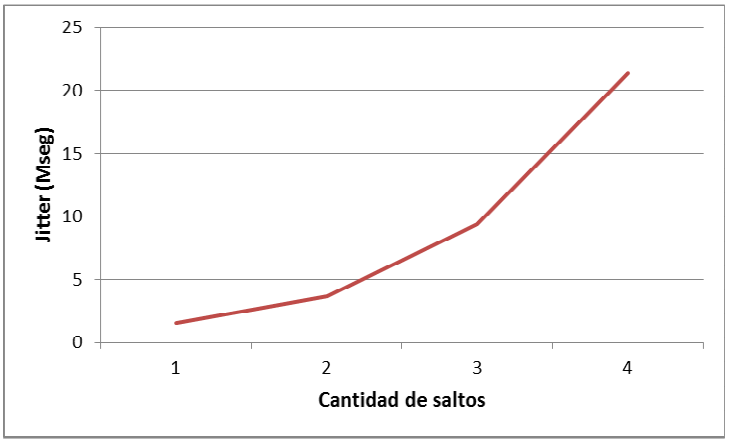


Fig. 3. Evaluación de Jitter en OLSR

La Figura 4 muestra los resultados de Probabilidad de error usando OLSR. Se observa que la pérdida de paquetes crece con el número de saltos y se vuelve significativa a partir de 3 saltos.

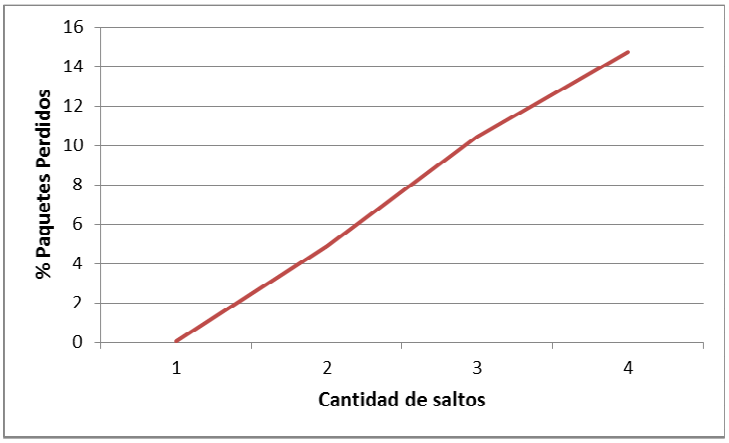


Fig. 4. Evaluación de la Probabilidad de Error en OLSR

La Figura 5 muestra los resultados obtenidos de las pruebas con TCP y UDP. En ambos casos el comportamiento es similar. El ancho de banda sufre un decrecimiento a medida que se incrementa el número de saltos.

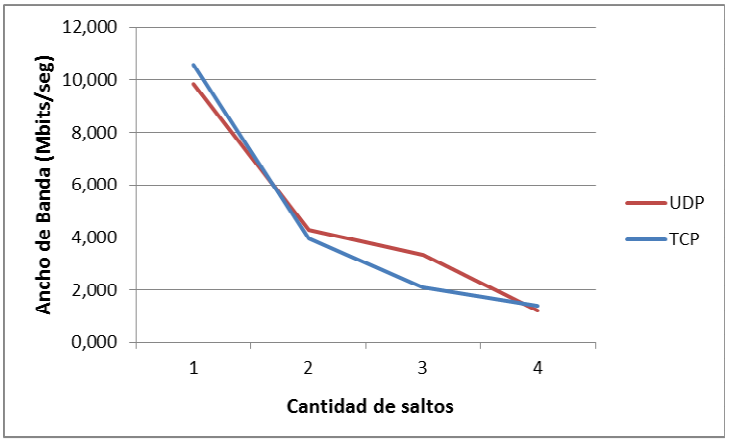


Fig. 5. Evaluación del ancho de banda en OLSR con UDP y TCP

Además, utilizando el mismo escenario, se realizaron pruebas para medir el tiempo de recuperación del protocolo cuando cae un nodo de la red y para medir el tiempo que tarda el protocolo en entrar en funcionamiento cuando se incorpora un nuevo nodo a la red.

Para medir el tiempo de recuperación cuando cae un nodo de la red, se procedió al apagado del nodo 9 y se midió cuánto tiempo tardaba la red mallada en encontrar un camino alternativo. El promedio obtenido fue de 25,63 segundos.

Para medir el tiempo de arranque, se apagó el nodo 14, luego se volvió a arrancar este nodo registrando la hora de encendido. Este procedimiento se realizó ejecutando un script para evaluar cuantos segundos demoraba en re arrancar. Se tomó la hora de la ejecución del primer ping inalcanzable y luego la hora del primer 0% paquetes perdidos. El promedio obtenido fue de 42,46 segundos.

Para complementar las pruebas con servicios reales se montaron sobre la red mesh, una central telefónica IP Elastix con cinco teléfonos internos: tres internos utilizando un ATA (Linksys phone adapter PAP2-NA) y dos por medio de software cliente de centrales IP. También se montó una cámara IP sobre uno de los nodos más alejados. En estas pruebas (video y voz sobre IP) sobre la red se pudo visualizar un desempeño aceptable de la misma para dichos servicios aunque la aplicación de video no responde eficientemente a cambios bruscos.

5 Conclusiones

En este trabajo se presentó un estudio sobre el rendimiento del protocolo OLSR. Se presentaron los resultados empíricos de varias pruebas utilizando el mismo escenario. El escenario que se presenta es una red de mundo real (no de laboratorio) con pruebas reales y

no simulaciones. Cuando se trata con este tipo de entornos, los experimentos son cada vez más difíciles de repetir en forma exacta al anterior.

Por lo observado se confirma que el rendimiento de la red decrece con el número de saltos. Su valor, que en nuestro caso es bajo para cuatro saltos, depende mucho de la ubicación y la conectividad entre dispositivos, como así también de la actividad radioeléctrica circundante.

A la luz de los resultados hay algunos descubrimientos interesantes: a) OpenWrt es un software perfectamente válido que puede ser utilizado en una gran variedad de dispositivos y b) su configuración para utilizarlo con protocolo OLSR es sencilla de realizar y no presenta problemas de funcionamiento con dicho protocolo.

Ex-profeso se utilizó hardware variado para demostrar que se puede realizar una mesh con los dispositivos disponibles en mercado (como los TP-LINK) e incluso algunos más antiguos, como es el caso de los Linksys WRT54GL. De todas maneras en el relevamiento de redes existentes realizado y en el grado de desarrollo de los firmware, se pudo observar que los equipos más utilizados son las distintas versiones de Nanostation de la marca Ubiquiti.

Las pruebas con servicios reales sobre la red no tienen rigor de investigación y fueron hechas a los efectos de visualizar en forma sencilla el comportamiento de la red y la validez de la provisión de estos servicios sobre la misma.

Cabe aclarar que por tratarse de una red montada sobre un escenario real y no de laboratorio hemos tenido que escoger adecuadamente los horarios de realización de las pruebas dado que las otras redes inalámbricas instaladas en el edificio y la circulación de personas tienen una marcada influencia en el funcionamiento de la red mallada.

Bibliografía

1. Akyildiz, I., Wang, X., Wang, W.: Wireless mesh networks: a survey, In Computer Networks. Vol. 47. No.4 pp. 455--487 (2005)
2. Acuña Martínez, D., Roncallo Kelsey, R.: Redes inalámbricas enmalladas metropolitanas. pp. 46--91 (2006)
3. http://www.olsr.org/ Consultado el 02/05/2013
4. http://www.open-mesh.org/ Consultado el 08/03/2013
5. http://wiki.freifunk.net/Kategorie:Espanol / Consultado el 01/04/2013
6. https://commotionwireless.net/ Consultado el 01/04/2013
7. https://openwrt.org/ Consultado el 08/03/2013

NetworkDCQ: A Multi-platform Networking Framework For Mobile Applications

Federico Cristina[1], Sebastián Dapoto[1], Fernando G. Tinetti[1,2], Pablo Thomas[1], Patricia Pesado[1,2]

[1] Instituto de Investigación en Informática LIDI - Facultad de Informática
Universidad Nacional de La Plata - Argentina

[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires - Argentina

{fcristina, sdapoto, fernando, pthomas, ppesado}@lidi.info.unlp.edu.ar

Abstract. Currently, the number of mobile applications that require (wireless) connectivity is constantly increasing. The need for sharing information among mobile devices exists in many applications, and almost every data exchange between these devices involve the same requirements: a means for discovering other mobile devices in a wireless network, establishing logical connections, communicating application data, and gathering information related to the physical connection. This paper proposes an open source developer-oriented framework that acts as a network support layer for host discovery, data communication among devices, and quality of service characterization, which can be used for developing several types of applications and is proposed for different platforms, such as Android Java, J2SE, and J2ME.

Keywords: mobile devices, host discovery, communication, QoS, networking

1 Introduction

The middleware presented in this paper, called NetworkDCQ, is proposed bearing in mind the evolution of mobile devices as well as specific network requirements of mobile applications. The following subsections briefly explain three topics: trends in mobile devices, mobile network applications, and the initial development platform selected for (a proof-of-concept) implementation.

The remainder of this paper is organized as follows. The next section describes the proposed Application Program Interface (API). Afterwards, an architectural overview of the framework is given. The following section presents several applications which use NetworkDCQ. Finally, we describe the results and benefits of using the proposed framework and conclude with an outlook on future work.

1.1 Trends in Mobile Devices

The worldwide internet mobile traffic is expected to overtake the desktop internet traffic by 2014 [1], which means that more users will be accessing the Internet through their mobile phones than through their PCs. This phenomena has already been experienced in some countries, like China [2] or India [3, 4].

Currently, nearly 50% of recent device sales are mobile (smartphones, tablets) [5]. Mobile applications are tightly related with this trend. The increasing number of these devices in the last years has led to a revolution in terms of mobile application development and usage. Among all OS mobile systems, Android is by far the most deployed platform [4, 6], with 136 million units shipped and 75% market share in Q3 2012 [7], seconded by iOS and BlackBerry OS with 14.9% and 7.7% market share respectively. Additionally, Android has a large community of developers writing applications that extend the standard functionality of the devices. Google play has hit the 25 billion-download mark by September 2012 [8].

1.2 Mobile Network Applications

Although there is a large number of standalone mobile applications (which require no connectivity at all), a currently increasing trend in mobile environments is the development of applications in which several devices on a network share real time information. These applications rely on some sort of connectivity support in order to achieve the proper interaction among devices. This support can be grouped into three main categories, or services: a) *Host discovery*, a mean for searching other reachable devices ready to communicate in a network, b) *Data communication,* a service for handling the specific exchange of information between devices, and c) *Quality of service,* a monitoring service that provides QoS related information.

Since these services are application-independent, a framework can be implemented in order to support specific aids, simplifying the network-related aspects to the developer. The main purpose of the proposed infrastructure is to meet these service's requirements. The features provided by *NetworkDCQ* allow several types of implementations with different network configurations, such as a typical client/server architecture or a centralized/decentralized peer-to-peer solution.

Even though there are several mobile development frameworks [9, 10], none of them proposes an open source, multi-platform solution that presents the features proposed in this paper. Some of these frameworks refer to *networking features* as simply retrieve wireless connection information, but no additional functionality is supported. Other frameworks cover these features, but as a part of a complete paid solution for mobile-apps development. The most representative examples are PhoneGap [11], Unity3D [12], Titanium [13] and Corona [14].

1.3 Development Platform

The reason for choosing Android as the primary development target for the proposed framework is based on its widespread use and popularity (as previously explained).

However, two additional benefits should be mentioned. First, it is an open source software released under the Apache License. This allowed several non-official versions such as Android for x86, ARM, and MIPS architectures. Some examples given in the present paper were tested on these versions running in a Virtual Machine, without the need for real devices. Second, Android Java is functionally much richer than J2ME. Actually, the similarities with J2SE API (Application Programming Interface) led to the Oracle vs Google lawsuit [15]. As will be shown, this is a considerable advantage due the compatibility between both languages in matters of network communication. This means that the proposed API can be referenced from both types of Java projects. Given that one of the purposes of the framework is to achieve multi-platform compatibility, a J2ME version is also being developed, allowing interoperability between the other platforms.

2 Proposed and developed API

The main goal is having a minimal (yet useful) communication-related software infrastructure so that different mobile devices can be programmed. The focus is on the Java language since it is (by definition) cross platform. Even when currently development platforms tend to be very different, it is possible to use Java in almost all of them. While the first problem to be solved is programmability, other issues such as interoperation are left open for future release/development. This section will present the main classes and interfaces of the framework from an application developer point of view. Based on the previous analysis, and the types of interaction required among hosts, the highest level of the API is directly focused on application data communication (Application Support) and the lowest level is divided into three main parts, as shown in Fig. 1:

- HostDiscovery, for handling the information related to hosts that are ready to communicate to/from each device. As its name suggests, HostDiscovery services/operations include searching for hosts and/or hosts status.
- NetworkCommunication, for handling the specific exchange of information between applications. Basically, NetworkCommunication should include the necessary send and receive services/operations for applications.
- QoSMonitor, for providing the user and/or programmer the necessary information on signal quality as well as performance indexes such as startup time (latency) and available network bandwidth.

The initial aim for each part is to achieve a very simple interface for the user, simplifying the API usage as well device programmability. As a general concept, the framework is designed to support different implementations for each of the services (Discovery, Communication, and QoS). Through an *Abstract factory* pattern [16], the user can specify which implementation should be used in each case. The details explained in this section go beyond any implementation, covering the issues at a higher level of abstraction.

2.1 Application Data, Producer, Consumer

Generally, the framework will require a data producer, a data consumer, and the data itself to be transferred among hosts. The three will be instances of user-developed classes which extend/implement a specific class/interface. Based on Inversion of Control [17, 18], these instances will be passed to the framework as arguments. Specific methods of the instances will be called from the framework in order to generate new data, process incoming data, handle a new host in the network, etc.

The base class for the application-level data is the abstract class *NetworkApplicationData*. This class will be the superclass for any information to be sent/received through the *NetworkCommunication* services. Currently, the only information contained in this class is a reference to the source host (the one that originates the message). Subclasses must augment the data structure as needed, and any data type/object can be used as long as it implements the *Serializable* interface.

The producer class is in charge of generating the updated local information to be sent to the other hosts. This class must implement the *NetworkApplicationDataProducer* interface. This interface only requires the method *produceNetworkApplicationData* to be implemented, which returns an instance of a subclass of *NetworkApplicationData* with the actual data. This method will be called periodically if the periodic Broadcast feature from the *NetworkCommunication* service is active. The period is given by the user in milliseconds, also provided by the API. If this feature is not desired, then there is no real need for this class to be implemented. However, it is advisable to centralize the creation of data in a specific class. In this case, calls to *produceNetworkApplicationData* method will have to be done manually from some application-level class when required.

The consumer handles every type of incoming information, mainly related to application data from other hosts as well as notifications of arrivals and departures of hosts to/from the network. Every time a new message arrives, the framework will invoke the *newData* method so that the application can act accordingly. A *NetworkApplicationData* object is received as a parameter, containing the actual data. The consumer will have to cast this object to the corresponding application-level data type. When the *HostDiscovery* service identifies some network change related to hosts, the corresponding method will be called. This allows applications to behave in a specific way in this cases. Thus *newHost* or *byeHost* methods will be called when there is a new host in the network or when a host leaves the network respectively.

2.2 Host Discovery

As mentioned above, this service is responsible of searching for new hosts in the network as well as exchange host status periodically. The status of a host is simply an online/offline flag in order to know if the host is ready to receive information at a certain moment. The discovery service can be started simply by invoking the *startDiscovery* method. This will make the framework to look/listen for/to new hosts, calling a specific method each time a host joins or leaves the network. When the service is not needed anymore, the *stopDiscovery* method can be invoked. This implies neither sending local status nor receiving other hosts status anymore.

The periodicity a host sends its status can be set depending on the application requirements. Making available *stopDiscovery* as well as the periodicity value to the programmer is necessary in order to have control on energy and communication overhead/usage. The current list of hosts which are part of the network can be accessed through the *otherHosts* collection so that at any time, the application would be able to search for specific hosts available and the total number of hosts with which could exchange information.

2.3 Network Communication

Network communication services (provided by *NetworkCommunication*) allow hosts to exchange application-level data in different ways, depending on the specific needs of the application being developed. Client/server, broadcast, and Producer/Consumer communication models are available to the applications. In order to establish an application-level communication with other hosts, the *startService* method must be started. Once started, the service waits for incoming connections from other hosts. A host can establish a connection to another host through the *connectToServerHost* method. An established connection will be used for sending and receiving the application-level data. When a message is received, a *Consumer* will be able to process the incoming information.

Sending a message simply implies specifying the target host and the data to be sent (using *NetworkApplicationData*, as mentioned above), through the *sendMessage* method. Additionally, a host might need to send information to every online host in the network calling the *sendMessageToAllHosts* method. When the service is not needed anymore, the *stopService* should be called. This will close all currently established connections.

Also, the framework is able to handle sending data to all hosts periodically. In this case, NetworkDCQ will require in each sending the updated local information. A *Producer* will have to generate this information. This feature is available by calling the *startBroadcast* method and is useful in cases when a constant exchange of data among hosts is needed at regular intervals, for instance in a network game. The application-level periodic data broadcast can be stopped by simply invoking *stopBroadcast* method. The periodicity a host sends data can be set depending on the application requirements.

2.4 QoS Monitor

A useful set of services is currently being defined, so that each application will be able to decide if it is possible to run under the current network bandwidth, signal strength, etc. At the lowest level of abstraction, an application should be able to ask for the current startup and available bandwidth, so that it will be possible to model the time required to send a message of n data items.

Also, some of these performance indexes would depend on wi-fi signal strength, so it would be useful to provide the application with the current signal strength as well as some previous values so that the tendency would be able to be estimated. From a

higher level of abstraction, a method such as *calculateMPS* for an estimation of the number of application-data messages per second would be able to be exchanged, and it would aggregate some low level information, along with the specific application data to be communicated periodically. Although an initial API is proposed, this service is currently under development and unavailable to user applications.

3 NetworkDCQ proposed architecture

This section will discuss in detail the implementation aspects of the proposed architecture. As mentioned before, the framework supports different implementations for each low level service. Currently, an *UDPDiscovery* and *TCPCommunication* was developed for *HostDiscovery* and *NetworkCommunication* services respectively, and *QoSMonitor* is under development. Fig. 1 shows the most relevant details on each layer, which will be explained in the following subsections (excepting QoSMonitor).

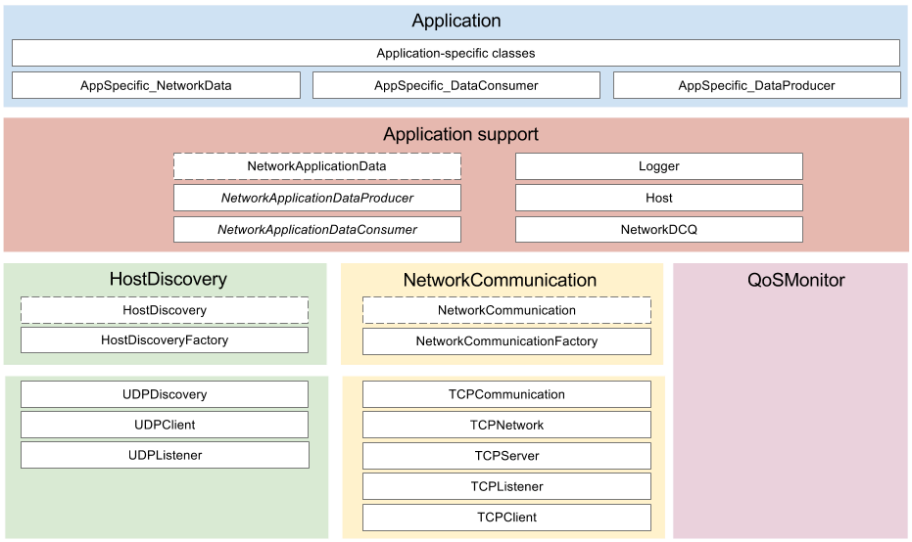


Fig. 1. Detailed Architecture of the Framework

In Fig. 1, abstract classes are identified with dotted lines, and interfaces are those in italic font. The current implementation of the project can be found at [19] hence the description in this section will be far from explaining the code (or code details), which can be downloaded, used, etc. Section 4 will explain in detail (via specific examples) the step-by-step guide in order to configure and use every feature of the framework.

3.1 Application support

This layer involves additional classes which are referenced along several parts of the

framework. For instance, *NetworkApplicationDataConsumer* is related with Discovery and Communication services. Host instances exist in Discovery, but they are also used in Communication. A special class in this layer is *NetworkDCQ*, which is explained in detail in the next subsection.

3.1.1 NetworkDCQ

This class is the framework main entry point, and has two main static methods. Method *configureStartup* allows the developer to specify the Producer and Consumer instances. Method *doStartup* is the one in charge of starting each service or feature (discovery, communication, broadcast), since they can be started independently. It is expected that *configureStartup* is called before any usage of the framework and method *doStartup* identifies the point from which the application would start using every framework service (discovering hosts, establishing communication/s, etc.).

3.2 UDPDiscovery

UDPDiscovery is the implementation of *HostDiscovery*, extending its abstract class. As such, it implements *startDiscovery* and *stopDiscovery* methods. When the discovery service is started, the *UDPDiscovery* spawns two threads: *UDPListener* and *UDPClient* as shown in Fig. 2a. The former first joins the network group via a *MulticastSocket*, and then waits for incoming host status updates from other hosts. The latter periodically sends multicast packets with its local host status.

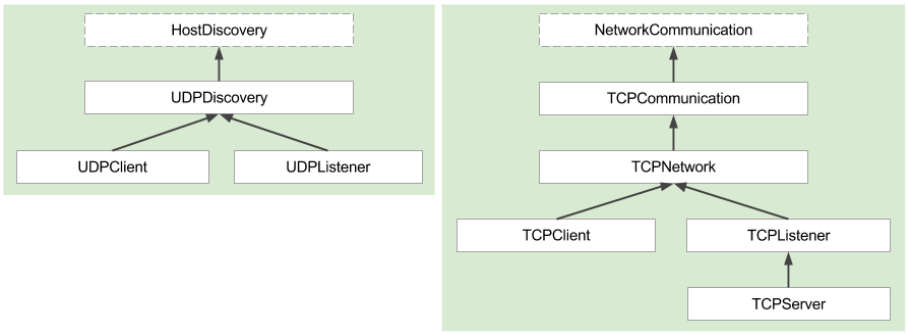


Fig. 2. a) *UDPDiscovery* Hierarchy, b) *TCPCommunication* Hierarchy

UDPDiscovery has an additional responsibility, which is to check for hosts that leave the network without giving the proper signal. This is achieved by a connection timeout validation, i.e. by checking - for each remote host - the timestamp of the last received status update. If the lapse of time exceeds a predefined threshold, then the host is removed from *otherHosts* list and *byeHost* method is invoked. This validation is executed periodically.

3.3 TCPCommunication

TCPCommunication is the implementation of *NetworkCommunication*, extending its abstract class. This service will spawn several threads, depending on the framework configuration. The following is a brief explanation of the methods discussed above and taking into account the details shown in Fig. 2b.

Method *startService* will spawn a *TCPListener*, in charge of listening for new TCP connections from other hosts. For each new connection, this class will spawn a new *TCPServer* thread, which is in charge of receiving *NetworkApplicationData* objects from a specific host.

Method *startBroadcast* will spawn a *TCPCommunication* thread, which will periodically send a *NetworkApplicationData* object (relying on the configured *NetworkApplicationDataProducer* that generates the data), using the *sendMessageToAllHosts* method. This last method simply iterates the *HostDiscovery.otherHosts* collection, and calls *sendMessage* method in each case.

TCPCommunication has a pool of *TCPClient* objects (the ones in charge of writing data through a socket), one for each host. Method *connectToServerHost* instantiates a new *TCPClient* when invoked and will keep it in the pool for later use. Every time a message is sent to a host, *TCPCommunication* first retrieves the corresponding connection with that host, avoiding having to reconnect continuously.

4 Examples

In this section three different examples will be discussed, in which the network requirements for each application differs considerably. The first one is a competitive multiplayer Asteroids-like game (referred to as Asteriods, from now on) and the second one is a two players Tic-Tac-Toe game, both currently running in Android. The third example is a simple chat application implemented both in Android and J2ME in order to show multi-platform communication.

In each case, sample code will be given in order to highlight the most relevant details related to networking. The complete code of the first two examples can be found at [20] and [21] respectively. Also, these projects are *completely* built on top of the NetworkDCQ project [19], i.e. there is no access to other Host Discovery and Communication services beyond those provided by the NetworkDCQ framework. For the third example, the J2ME version of the chat application is built on top of the J2ME version of the NetworkDCQ project [22].

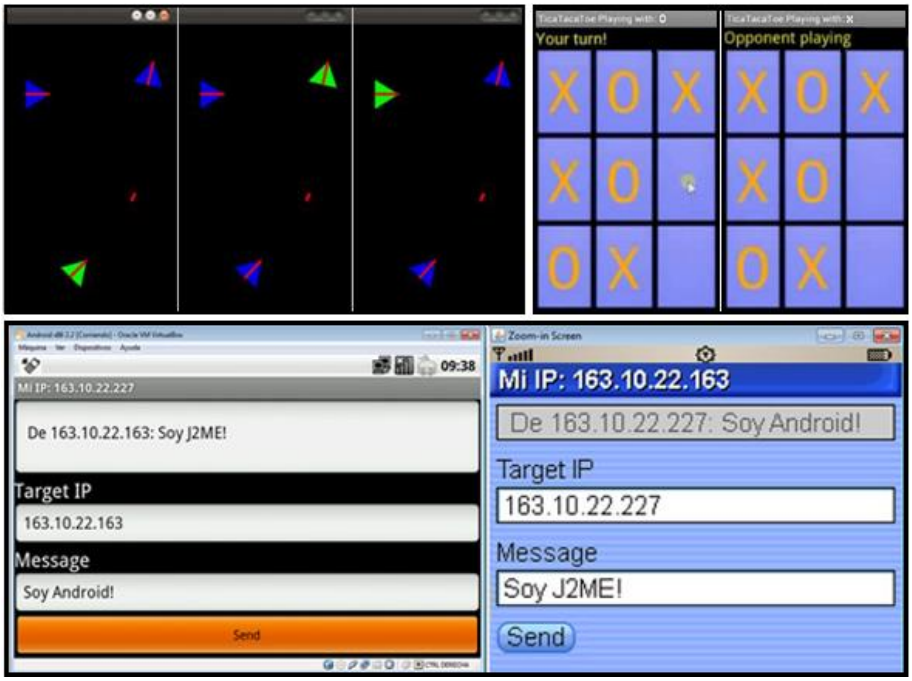


Fig. 3. a) Asteroids running on three Android x86 v2.2 virtual machines, b) Tic-Tac-Toe running on two Samsung Galaxy SII mobile devices with Android 4.0.3, c) Chat application running on Android x86 (left) and J2ME Emulator (right).

4.1 Asteroids

Multiplayer Asteroids is a very simple game, in which a ship (controlled by a user) must destroy enemy ships firing laser shots. Every ship corresponds to a user in a host (i.e. mobile device, tablet, etc.) in the network, as shown in Fig. 3a. The local ship will be rendered in green and remote ships will be rendered in blue. An example video of the game can be found at [23], where it is also shown that the entire example is run on virtual machines with Android.

Although very basic, the application is representative in terms of CPU and network usage of a class of game applications: the game must continuously update its local model, share local information among all hosts, receive and update remote hosts information, and render the corresponding graphics. Considering an update rate equivalent to 30 frames per second, the network consumption is considerably high and grows proportionally to the number of players. Furthermore, the game uses the *Periodic Broadcast* feature from the *Communication* service.

The data defined to be sent/received through the network includes ship position and heading, as well as shots position and heading that the ship shoots when the user triggers the fire action. The producer has a unique and reusable *AsteroidsNetworkApplicationData* instance (in order to avoid continuous Garbage Collector calls), which is filled in new data every time is needed with its current

values according to the model changes. The Consumer is the place where remote ships information is updated with the received data. A cast to *AsteroidsNetworkApplicationData* is needed in order to retrieve the members in the instance (ship heading, position, etc.). The last step simply requires setting the corresponding application-level instances of *Producer* and *Consumer* of the framework, and starting the Discovery, Communication and Broadcast services.

4.2 Tic-Tac-Toe

Tic-Tac-Toe has been selected as a representative example of a completely different type of application, compared to the Asteriods game, since Tic-Tac-Toe is a two-players game, turn-based and there is no need for a continuous sending of information, specific events (players taking turns) trigger communications.

Fig. 3b shows a running example of the game on two Samsung Galaxy devices with Android 4.0.3, and an example video of the game running on a virtual machine and a Samsung Galaxy can be found at [24]. While the Tic-Tac-Toe game impose a very different usage of the network during the game (turns, non-periodic messages, etc.) as compared to the Asteroids game, other service requirement such as those related to host Discovery remain the same.

The data structure for this application is very simple: an action value representing the possible states of the game: a) resolve who will start the game, b) set a cell with an X or an O - in this case a position value is also needed, or c) restart the game. Since there is no need for a periodic update of local host information, no *Producer* has to be implemented. The Consumer is the place where each remote action is replicated locally (e.g.: the other player placed an X in cell 7). A cast to an application-level data type is needed in order to retrieve the members in the instance (action and position if needed). As explained previously, the sending of information is not performed periodically. The application sends a message to the other host each time an action event occurs (e.g. when the user clicks in one of the nine cells).

The application access the *Communication* service through the static method *NetworkDCQ.getCommunication* in order to use the *sendMessage* method. The other host is retrieved by accessing the *HostDiscovery* static member *otherHosts*. The final step is starting the required services. In this case, the Producer and the Broadcast service will not be started.

4.3 Multi-platform chat application

A simple chat application has been selected in order to show multi-platform networking capability, requiring only the NetworkDCQ communication features. By simply specifying an IP address and a message, the chat-app sends the corresponding text to the target host, the which shows its content on the display. Fig. 3c shows the achieved interaction among two virtual devices, one running the application on Android, and the other running on J2ME.

The biggest problem in this case is the serialization-deserialization issue. Each platform implements (if it does) a specific serialization method, which can or cannot be compatible with the other platforms. In order to solve this problem, NetworkDCQ defines a *NetworkSerializable* interface, containing the definition for the *networkSerialize* and *networkDeserialize* methods. Applications must contain a class which implements this interface in a consistent way on each platform. At run time, NetworkDCQ then delegates the serialization-deserialization work to these classes.

5 Conclusions and Further Work

The paper presented a framework for handling network-related issues in the development of applications running on mobile devices, such as host discovery, data communication and broadcasting; designed to support different implementations for each of these services, gaining flexibility, and versatility. Its main goal is to fill a gap in the mobile development frameworks area, where currently there is no open source, multi-platform solution with the features proposed in this paper.

The proposed API and reference implementation is actually useful for several types of applications, network requirements, and configurations. The examples shown in the previous section cover applications with a wide variety of network-related requirements like continuous data broadcasting and event driven communication.

Using Android as a general development platform allowed an immediate integration with J2SE applications. Additionally, specific interaction problems with other platforms where solved by defining the corresponding interfaces and development methodologies, allowing communication with platforms such as J2ME.

As explained previously, the QoS service is still in development. Completing this feature is a short-term objective. Implementing the complete set of features for iOS, Windows Mobile, and BlackBerry 10 are mid to long-term objectives.

References

1. Morgan Stanley. The Mobile Internet Report, 1st edition. (2009)
2. China Internet Network Information. China Internet Development Statistics Report. (2012).
3. Mobile vs Desktop Internet Traffic Report from Oct 2011 to Oct 2012. http://gs.statcounter.com/#mobile vs desktop-IN-monthly-201110-201210.
4. Meeker, M. D10 Conference. Internet Trends. (2012), http://www.kpcb.com/insights/2012-internet-trends.
5. Asymco. The Rise and Fall of Personal Computing (2012), http://www.asymco.com/2012/01/17/the-rise-and-fall-of-personal-computing/.
6. Gartner, Inc. Nov.2012 Press Release, http://www.gartner.com/it/page.jsp?id=2237315.
7. IDC. Nov.2012 Press Release, https://www.idc.com/getdoc.jsp?containerId=prUS23771812.
8. Google, Inc. Google Official Blog (2012), http://officialandroid.blogspot.com.ar/2012/09/google-play-hits-25-billion-downloads.html
9. Markus Falk. Mobile Frameworks Comparison Chart, http://www.markus-falk.com/mobile-frameworks-comparison-chart/
10. Digital Possibilities. Mobile Development Frameworks Overview

http://digital-possibilities.com/mobile-development-frameworks-overview/
11. PhoneGap, http://phonegap.com/
12. Unity3D, http://unity3d.com/
13. Titanium, http://www.appcelerator.com/platform/titanium-platform/
14. Corona, http://www.coronalabs.com/products/corona-sdk/
15. Reuters. Oracle sues Google over Android (2012), http://www.reuters.com/article/2010/08/13/us-google-oracle-android-lawsuit-idUKTRE67B5G720100813
16. Gamma, E. Design Patterns: Elements of Reusable Object-Oriented Software (1994).
17. Martin, R. C. The Dependency Inversion Principle. (1996), http://www.objectmentor.com/resources/articles/dip.pdf
18. Fowler, M. Inversion of Control Containers and the Dependency Injection Pattern, http://martinfowler.com/articles/injection.html
19. NetworkDCQ for Android Project, https://code.google.com/p/networkdcq/
20. Asteroids for Android Project, http://code.google.com/p/asteroidsa/
21. Tic-Tac-Toe for Android Project, http://code.google.com/p/ticatacatoe/
22. NetworkDCQ for J2ME Project, https://code.google.com/p/networkdcq-j2me/
23. Asteroids for Android Example Video, http://www.youtube.com/watch?v=HiRTk8daqi4
24. Tic-Tac-Toe for Android example video, http://www.youtube.com/watch?v=mrf01putSec

Estimación de “H” con transformada ondita

Reinaldo Scappini[1], Luis Marrone[2],

[1] UTN Facultad Regional Resistencia, calle French 414 Resistencia, Rep. Argentina
rscappini@gmail.com

[2] LINTI , Facultad de Informática-UNLP, cqalle 50 y 120 La Plata, Rep. Argentina
lmarrone@linti.unlp.edu.ar

Abstract. El análisis de tráfico se ha convertido en un proceso fundamental a la hora de evaluar la performance de una red. También se ha tornado crítico en la actualidad por la presencia de componentes auto-similares en él. Esta componente cambia el paradigma del modelo de tráfico utilizado hasta hace unos pocos años con serias dificultades analíticas; por lo menos comparándolos con los utilizados hasta el momento. Un parámetro clave en este nuevo modelo es el parámetro “H” o de “Hurst” por lo que importa una correcta detección y estimación. Presentamos con esa motivación los resultados obtenidos de la aplicación de un “script” basado en la transformada ondita o “wavelets”.

Keywords: tráfico, autosimilaridad, onditas, parámetro H, QoS, performance, modelos

1 Introducción

Una actividad fundamental en la evaluación de performance y diseño de las redes telemáticas, es el análisis del tráfico que transportan, materializado en parámetros tales como, tiempos de arribo, longitud de los mensajes, tiempos de transmisión, comportamiento en diferentes escalas de tiempo, etc. Este conocimiento permite optimizar los recursos de las redes, y también posibilita, que los servicios ofrecidos, cuenten con la calidad requerida. El logro este objetivo, es un área activa de estudio e investigación.

Promediando el año 1990, estudios realizados sobre muestras de tráfico tomadas de redes en funcionamiento, han demostrado en forma inequívoca que el tráfico tiene propiedades autosimilares, esto es, la existencia de patrones estadísticos o comportamientos que se repiten a diferentes escalas de tiempo. Un tráfico con características autosimilares, afecta en forma negativa el desempeño de la red. Se puede observar que el retardo promedio de los mensajes resulta mucho mayor que lo previsto por el análisis de colas tradicional.

Esto representa un inconveniente por partida doble. Una peor performance y la imposibilidad de un tratamiento analítico completo.

En este escenario, con tráfico originado en diversas fuentes, con sus respectivas características y particularidades, abordar un estudio para cuantificar o medir de manera apropiada la demanda que los usuarios imponen sobre los recursos de una red, requiere del uso de modelos que representen de una manera eficaz y eficiente un comportamiento compatible con las características observables en el tráfico real.

Surgen entonces, dos desafíos de importancia central:

- El desarrollo de modelos generales que abarquen las principales características del tráfico a estudiar.
- El desarrollo de aplicaciones, que utilizando esos modelos, permitan obtener conclusiones válidas.

El éxito de los modelos autosimilares radica en su capacidad de capturar las complejas dependencias que muestra el tráfico a distintas escalas de tiempo mediante el uso de pocos parámetros, en particular el parámetro de Hurst "H".

Dada la importancia del mismo en la caracterización del tráfico, es necesaria su correcta detección y estimación.

Si bien este trabajo muestra en forma resumida las ventajas de un método particular, todos los detalles y desarrollos teóricos se pueden encontrar en un trabajo previo [1] donde se analizaron en mayor profundidad.

En particular brindaremos los resultados obtenidos aplicando "LDestimate" [2], una script para la estimación del parámetro "H" implementada para el software Matlab® y basada en la transformada wavelet u ondita.

2 ¿Por qué utilizar la transformada Ondita ("Wavelets")?

En la lectura de diversos estudios e investigaciones realizados en los últimos años sobre el tráfico autosimilar en las redes telemáticas, se evidencian las ventajas que aportan los métodos basados en las wavelets u onditas, atendiendo a criterios de validez, confianza estadística y eficiencia computacional. En consecuencia, la utilización de las wavelets se ha convertido en una útil y eficaz herramienta para tareas de análisis, detección, estimación, modelado y simulación en el ámbito del tráfico autosimilar.

Como se menciona en las referencias bibliográficas [3] Pág. 84, y [4] Pág. 23; entre las ventajas del estudio del tráfico autosimilar por medio de wavelets u onditas podemos mencionar:

- La wavelets u onditas ofrecen un marco teórico que se puede aplicar tanto a procesos autosimilares, procesos LRD (Long Range Dependence, o dependencia de rango largo), trazas muéstrales etc. Pudiendo hacer un análisis en el dominio de la escala, de forma que se adapta "naturalmente" a las necesidades de poder estudiar un comportamiento en este dominio.
- Permite la división controlada de un proceso madre de variabilidad extrema, en subprocesos a diferentes escalas tornando manejable su comportamiento, aprovechando la independencia de los subprocesos obtenidos, se pueden emplear herramientas de la estadística clásica sobre las secuencias de los coeficientes wavelets, y de esta forma poder diseñar estimadores simples y eficientes.
- Los bancos de filtros de análisis y síntesis, proporcionan una forma computacionalmente eficiente de llevar a cabo tareas de análisis y síntesis de procesos autosimilares.

2.1 Relación entre las Wavelets y los procesos Hss, Hsssi y LRD

Hss es un proceso puramente autosimilar, Hsssi es autosimilar con incrementos estacionarios, LRD es cuando presenta dependencia de rango largo.

Si bien no es motivo de este trabajo, el estudio de la transformada Waveletet u Ondita, por cuestiones de contexto es importante señalar que existe un cuerpo teórico llamado análisis multiresolución (MRA) que propone la existencia de una función llamada Wavelet madre y cuyo producto interno con la señal $"S"$ representada por el procesos estocástico $X(t)$, da como resultado dos conjuntos de coeficientes llamados aproximación $a_s(j,k)$ y detalles $d_s(j,k)$ respectivamente, que preservan las características de la señal original y permiten su estudio a distintos niveles de escalas frecuenciales y temporales. El parámetro j representa el nivel de escala también denominado octava y el parámetro k la traslación o desplazamiento temporal.

Si la señal $X(t)$ es proceso un proceso estocástico que presenta un fenómeno de escala representado por el exponente $"\alpha"$, los coeficientes correspondientes a su transformada wavelet, tendrán las siguientes características:

El conjunto $\{d_X(j,k), j=1,2,...J, k\in \square\}$, es un proceso estacionario para cada octava *j*, si el Nº de momentos desvanecientes de la wavelet madre ψ_0, es $N \geq \frac{\alpha-1}{2}$. La varianza del conjunto $d_X(j,k)$ reproduce el comportamiento de escala subyacente, dentro de un rango de octavas $j_1 \leq j \leq j_2$. Dado que el valor medio de la wavelet es cero por la condición de admisibilidad, el segundo momento de $d_X(j,k)$ es proporcional a $2^{j\alpha}$, donde $j_1, j_2\ y\ \alpha$, dependen del tipo de fenómeno de escala que exhiba el proceso original $X(t)$. Se cumplen entonces, las siguientes relaciones entre estos tres parámetros de la siguiente ecuación:

$$E\left[d_X(j,k)^2\right] \approx 2^{j\alpha} \qquad \textbf{(1)}$$

Si $X(t)\ es\ Hsssi \rightarrow\ \alpha = 2H+1,\ y\ -\infty < j < \infty$

Si $X(t)$, presenta *LRD*, $\alpha = 2H-1$; $j_2 = \infty$, y j_1 debe identificarse en función de los datos obtenidos en el análisis.

En caso de que el proceso obedezca una ley de potencias, pero en un determinado rango de frecuencias, $f1 \leq f \leq f2$, (a este tipo de procesos se los denomina genéricamente procesos $1/f$), γ corresponde a la ley de potencias expresada y el rango de escalas (j_1, j_2), debe obtenerse partiendo de las frecuencias (f_1, f_2).

3 Estimación mediante la script LDestimate

La script LDestimate, a diferencia de otras herramientas utilizadas en la estimación de *"H"*, proporciona información extra acerca del contexto en el que se estima *"H"*, esto es, tiene funciones accesorias que nos permiten escoger con bastante certeza la octava donde se inicia la alineación descartando las regiones que producen sesgo en el resultado, proporcionando además un estadístico que nos indica la "bondad del ajuste", en función de los valores estudiados, y asegura que el rango escogido tenga una efectiva alineación, evidenciando el fenómeno de escala y no una simple aproximación promediada con eventuales desviaciones propias de la técnica de regresión lineal. En los otros métodos la estimación en forma general se hace tomando la máxima cantidad de datos y se pueden producir importantes desviaciones que no son tenidas en cuenta por carecer de estas funciones tales como mostrar el intervalo de confianza y la bondad del ajuste, se muestra a continuación los fundamentos del método y unos comentarios acerca de los parámetros involucrados

3.1 Diagrama Log-escala – Estimación de H

La gran ventaja estadística del análisis Wavelet en el dominio de las escalas se evidencia en la expresión:

$$E\left[d_X(j,k)d_X(j,k')\right] \approx \left|k-k'\right|^{\alpha-1-2N} \text{ ; a medida que } \left|k-k'\right| \to \infty \qquad (2)$$

Esto nos permite medir los promedios temporales y utilizarlos como estimaciones de los promedios estadísticos, y la falta de correlación entre los distintos coeficientes wavelets asegura que los estimadores temporales tengan una varianza pequeña.

Por otra parte la expresión $E\left[d_X(j,k)^2\right] = 2^{j(2H+1)}E\left[d_X(0,k)^2\right]$; puede tomarse como un estimador del espectro de potencia del proceso en las inmediaciones de la frecuencia correspondiente a la octava *j*, y como se demuestra en [1], sec. 3.3.4 y 3.7, se puede estimar la varianza del proceso $d_X(j,k)$, según:

$$\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j}\left|d_x(j,k)\right|^2 \qquad (3)$$

Donde n_j es el número de coeficientes en la octava *j*, y se observa que la varianza decrece conforme n_j aumenta, entonces la variable μ_j es una forma eficiente de representar en forma compacta el comportamiento de segundo orden del proceso estudiado $X_{(t)}$ en la octava *j*, y si se tiene en cuenta la expresión, los μ_j son prácticamente independientes entre sí generando un desacoplamiento del comportamiento de $X_{(t)}$ en las distintas octavas *j*, y dado que en el estimador la varianza decae hiperbólicamente en función de la octava *j*, se puede asumir que el

exponente de escala del proceso α, lo podemos estimar como una regresión lineal de $y_{(j)} \equiv \log_2(\mu_j)$, como una función de la octava *j*.

En general se puede afirmar que un proceso que presenta LRD tiene una densidad espectral que obedece a:

$$f(\nu) \approx \frac{cf}{|\nu|^{\alpha}}, cuando\ \nu \to 0 \quad (4)$$

Donde ν es la frecuencia; cf es una constante positiva, y $\alpha = 2H - 1$.

Es sabido que la densidad espectral o potencia del proceso es proporcional al segundo momento de la variable y con relación a lo expuesto en el punto 3.1 (eq.1, 2 y 3), se pueden establecer equivalencias en ambas expresiones y expresar lo siguiente:

$$log_2\left(E\left[d_X(j,k)^2\right]\right) = j\alpha + log_2(c) \quad (5)$$

Lo que nos lleva a la siguiente aproximación:

$$log_2(\mu_j) = j\alpha + log_2(c) \quad (6)$$

A la gráfica de esta recta de regresión (eq.6) acompañada de los correspondientes intervalos de confianza en cada punto calculado, se la conoce como Diagrama Log-escala y de la pendiente se puede extraer el valor de α y despejar el valor de *H*.

En realidad lo expuesto hasta aquí es el fundamento básico del método, donde la pendiente α, se puede estimar en la región del diagrama donde los puntos se alinean entre dos octavas j_1, j_2 , dado que es posible que no exista alineamiento en otras regiones.

Según las notas que acompañan esta script, el autor parte de la definición de LRD dada en términos del espectro de potencia $f(v) \approx cf(v)^{-\alpha}$ cuando $v \to 0$, donde v, es la frecuencia, α es el exponente de escala, que es adimensional, y cf es un coeficiente con dimensión de varianza y describe aspectos cuantitativos de la longitud o extensión del comportamiento LRD, y como ejemplo de la importancia de cf expresa que los intervalos de confianza de la estimación de la media de LRD son proporcionales a $\sqrt{cf}$. La script entrega una estimación de sendos parámetros α y cf junto a otros que toman importancia según el contexto como se muestra a continuación.

- El diagrama Log-escala, nos proporciona una gráfica con la bondad de la estimación en cada punto como función de j_1 (figura izquierda), lo que permite una mejor elección de las octavas $j_{1,} j_2$
- Diferentes tipos de escala son posibles, sin embargo, el procedimiento de análisis es el mismo en cada caso. En primer lugar, el diagrama de Logscale se genera, y

examina los datos para encontrar un punto de corte inferior de la escala j_1, y otro punto de corte superior j_2, en los que la alineación (línea recta) se observa. Estos puntos de corte deben ser experimentados para encontrar un rango que se ajuste a la regresión de los intervalos de confianza sobre el Diagrama Logscale (Los valores iniciales se deben especificar en la lista de argumentos de "LDestimate", pero estos se pueden cambiar interactivamente.).

- Para cada rango de alineación elegido, la función de los resultados de la estimación de la pendiente "alfa", toma valores reales. El valor de alfa, y el rango j_1, j_2, ayudará a determinar qué tipo de escala se presenta.
- Por conveniencia, alfa se transforma en valores de los parámetros relacionados, como el parámetro de Hurst *H*, o la dimensión fractal de la muestra (válida sólo si es gaussiana). El usuario debe determinar qué tipo de escala está presente.
- NOTA: en el caso de LRD, alfa es el parámetro pertinente directamente, sin embargo, a veces es reescrita como una «H», pero esta no es la H de auto-similitud estricta, es simplemente una convención para reescribir alfa de esta forma para procesos LRD.
- La experimentación con el número momentos desvanecientes N de la wavelet es necesario para: (a) asegurarse de que la onda detalles están bien definidas (con valores de *H* altos, serán necesarios valores más altos de N, N = 1 es suficiente para estudiar LRD), y (b) eliminar o disminuir la influencia de las tendencias deterministas, como ser tendencias lineales o variaciones de nivel medio, que pueden estar presentes. En ambos casos es conveniente un aumento de N, hasta logar un diagrama Logscale estable.
- Una estadística de bondad de ajuste $Q(j_1)$, basado en Chi-Cuadrado se emite para ayudar con la elección del rango de escala, y se representa en el título del gráfico correspondiente (Fig.1.), como se muestra a continuación:.

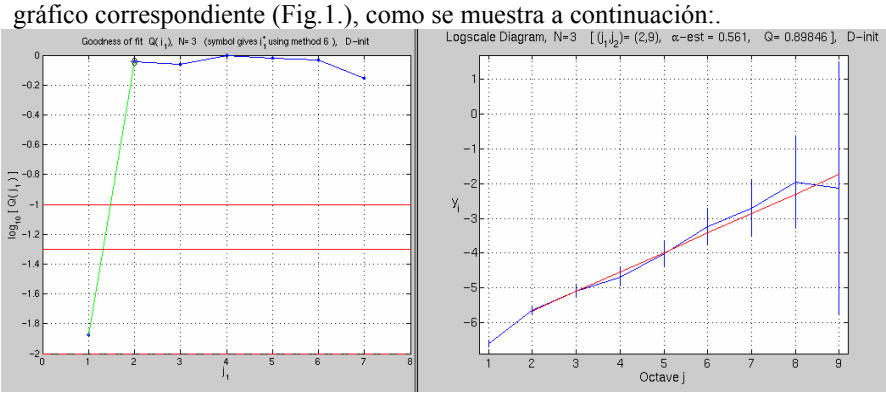


Fig. 1.

Es la probabilidad de observar que los datos, con las estimaciones de la varianza en cada escala realmente siga la forma de una recta para asegurar la efectividad de la regresión lineal. Un valor superior a 0,05 es aceptable y es aconsejable la

elección de $j_{(1)}$, a partir de donde se estabiliza el valor de $Q(j_1)$, La estimación visual de la región de alineación es difícil, cuando los intervalos de confianza se vuelven muy pequeños, en escalas pequeñas, en este caso la estadística $Q(j_1)$, es una mejor guía.

4 Utilizando LDestimate

La script puede utilizarse como una función con 7 argumentos de entrada con la siguiente sintaxis:

LDestimate(data,regu,j1,j2,discrete_init,calcj1,printout)

- data = Vector con los datos que se desean analizar (debe ser un vector fila).
- regu = N o número de momentos desvanecientes de la wavelet, este parámetro está relacionado con la regularidad de la estimación, a mayor valor mejoramos la bondad del ajuste Q, la variable regu está disponible desde 1 a 10. (sugiero empezar con 1 e ir aumentando para observar la variación de Q y el grafico de regresión respectivamente).
- j1 = octava de corte inferior debe ser ≥ 1.
- j2 = octava de corte superior, su valor máximo está relacionada con la longitud de los datos, pero se puede cambiar interactivamente durante la ejecución de la script, sugiero arrancar con el valor 2 y luego ir probando otros valores.
- discrete_init = con el valor 1 realiza la inicialización MRA para datos intrínsecamente discretos si el valor es cero, asume la de entrada de datos ya inicializado (es decir, ya está calculada la aproximación de secuencia o se utiliza un vector correspondiente a una aproximación).
- calcj1 = con el valor 1 realiza el cálculo de j1 optimo utilizando la script newchooseJ1 mencionada más arriba, si el valor es cero omite el cálculo. (sugiero dejarlo en 1).
- printout = con el valor 1 realiza los dos gráficos correspondientes a "Q" y la regresión del diagrama Log-escala. Con el valor cero no realiza los gráficos y solamente entrega los valores calculados.

5 Análisis realizado

A continuación y a manera de muestra de las posibilidades de estudio que brinda esta metodología, se analizan una serie de trazas que están disponibles para su uso en la investigación conforme las políticas de cada fuente (más detalle a continuación en las correspondientes descripciones).

El análisis se realiza sobre muestras tomadas de siete tipos de enlaces:

1. Ethernet 10 Mbits.
2. Ethernet 100 Mbits.
3. Ethernet 1 Gbits.
4. Ethernet 10 Gbits.
5. OC12
6. OC48
7. OC192

- Trazas Ethernet 10 Mb: Son las clásicas trazas utilizadas en muchísimos trabajos y que fueran utilizadas en el trabajo seminal de Leland et.al [5]; Si bien trabajo es de mucha antigüedad, es considerado como el punto de partida en este campo de análisis y es muy útil como referencia, pues casi todos los estudios comparativos realizados por distintos investigadores lo utilizan
- Trazas Ethernet 100 Mb: Las trazas pertenecen a la colección: "WIDE-TRANSIT 100 Megabit Ethernet Trace (Anonymized).". Se encuentran disponibles en [6].
- Trazas Gigabit Ethernet: Estas trazas corresponden a capturas realizadas por NLANR PMA, mediante una tarjerta Endace DAG4.2GE dual Gigabit Ethernet network. Las trazas encuentran disponibles en la página del proyecto en el link [7].
- Trazas 10 Gigabit Ethernet Cluster TeraGrid SDSC: Estas trazas se recogieron mediante un monitor del proyecto PMA [8] (Passive Measurement and Analysis)
- Trazas sobre Fibra Óptica: Tres grupos de trazas sobre fibra óptica, OC12 – OC48, y OC192; fueron facilitadas por el Proyecto CAIDA [9] (CAIDA: The Cooperative Association for Internet Data Analysis).

5.1 Procedimiento previo aplicado a las trazas

Debido al importante tamaño de los archivos de las trazas (típicamente del orden de los gigabytes), para poder procesarlos con PC's convencionales, a todas las trazas se las sometió al siguiente tratamiento: Se utiliza el software Wireshark [10], que es un conocido analizador de protocolos basado en tcpdump, que permite manipular trazas que utilicen formato compatible con el tcpdump y además cuenta con una utilidad de línea de comandos llamada Tshark [11], que resulta particularmente apropiada para esta tarea como se muestra a continuación:

- Se toma el primer millón de tramas de la traza y se lo convierte en un archivo con la extensión .pcap. La sintaxis del comando es: Tshark –r [nombre de la traza] –c 1000000 –w [nombre.pcap].
- Si se quiere trabajar con los tiempos entre arribos, creamos el archivo correspondiente, de la siguiente manera: Tshark –r [nombre.pcap] –e frame.time_delta –T fields > nombre.txt. Esto lo que hace es, leer el archivo .pcap que se creo con el millón de trazas, y lo filtra mediante el contenido del campo timestamp, del que a su vez establece la diferencia con la lectura anterior creando el valor tiempo entre arribos para cada trama y luego guarda el archivo en formato ASCII.
- Del mismo modo si se desea trabajar con la longitud en bytes de la trama, se utiliza el campo frame.len para el filtrado de la siguiente forma: Tshark –r [nombre.pcap]

–e frame.len –T fields > nombre.txt. Obteniendo de esta forma la salida en formato ASCII que es la más cómoda para poder utilizarla con el Matlab.

Aclarados todos los detalles que permitirán repetir los análisis aquí expuestos, a continuación se muestra el resumen de los resultados del análisis efectuado a estas trazas, con el primer millón de datos para cada una de ellas, los datos corresponden a tiempos entre arribos y se muestran en la siguiente tabla.

6 Resultados obtenidos

Se exhiben los resultados de la estimación del parámetro *H* realizada con la script de Darryl Veitch, tomando una muestra de cada traza, conforme se describe anteriormente. Se aclara que las condiciones iniciales para todas las estimaciones son las mismas, es decir inicialmente se fijan los parámetros como: N= 4 momentos desvanecientes, $j_1 = 2; j_2 = 20$ y los tres parámetros restantes igual a uno, esto significa que la script calculará en función de los resultados iniciales para cada caso lo siguiente: El mejor valor para j_1 como función del j_2 que se encuentre como óptimo. Tomando los valores sugeridos por la script para j_1 y j_2, se repite la estimación obteniendo los datos que se muestran en la tabla 1:

Parámetro / Traza	j_1	j_2	α	α [95%]	*H*	*H* [95%]
10 Mbit. pAug89.TL	9	16	0,642	[0.590, 0.693]	0,821	[0.795, 0.846]
10 Mbit. pOct89.TL	6	16	0,51	[0.494, 0.527]	0,755	[0.747, 0.763]
100 Mbit. Tokio (20070109080 0)	8	16	0,315	[0.280, 0.349]	0,657	[0.640, 0.675]
100 Mbit. Tokio (20070109120 0)	6	16	0,304	[0.288, 0.321]	0,652	[0.644, 0.661]
Gigabit-Ethernet (20040130-132000-0)	10	16	0,557	[0.479, 0.634]	0,778	[0.740, 0.817]
10 Gigabit Ethernet (20040212-130000-0)	10	14	0,508	[0.286, 0.731]	0,754	[0.643, 0.865]
ampath-oc12.20070109.dag0.20070109-0000.anon	10	16	0,693	[0.616, 0.771]	0,847	[0.808, 0.885]
ampath-oc12.20070109.dag0.20070109-1500.anon	12	16	0,303	[0.106, 0.499]	0,651	[0.553, 0.750]
OC48-20020814-103000-0-anon.pcap	5	16	0,195	[0.184, 0.207]	0,598	[0.592, 0.603]
OC48-20020814-115000-0-anon.pcap	4	16	0,175	[0.167, 0.183]	0,588	[0.584, 0.592]
equinix-chicago.dirA.20090115-060100.UTC.anon	9	16	0,461	[0.410, 0.512]	0,73	[0.705, 0.756]
equinix-chicago.dirB.20090115-060100.UTC.anon	8	16	0,282	[0.248, 0.317]	0,643	[0.624, 0.659]

Table 1.

6.1 Tabla Comparativa de Resultados Obtenidos con Distintos Métodos de Estimación de "*H*".

La tabla 2, muestra los resultados de la estimación de H, utilizando las aplicaciones desarrolladas en [1], junto al resultado obtenido mediante la script de Darryl Veitch.

Traza \ Metodo	1	2	3	4
10 Mbit. pAug89.TL	0,802	0,737	*0,821*	0,756
10 Mbit. pOct89.TL	0,844	0,839	*0,755*	0,785
100 Mbit. Tokio (200701090800)	0,675	0,663	*0,657*	0,667
100 Mbit. Tokio (200701091200)	0,696	0,686	*0,652*	0,625
Gigabit-Ethernet (20040130-132000-0)	0,69	0,729	*0,778*	0,682
10 Gigabit Ethernet (20040212-130000-0)	0,93	0,979	*0,754*	0,883
ampath-oc12.20070109.dag0.20070109-0000.anon	0,685	0,727	*0,847*	0,683
ampath-oc12.20070109.dag0.20070109-1500.anon	0,638	0,678	*0,651*	0,684
OC48-20020814-103000-0-anon.pcap	0,58	0,651	*0,598*	0,626
OC48-20020814-115000-0-anon.pcap	0,66	0,645	*0,588*	0,637
equinix-chicago.dirA.20090115-060100.UTC.anon	0,653	0,667	*0,73*	0,642
equinix-chicago.dirB.20090115-060100.UTC.anon	0,693	0,698	*0,643*	0,637

Table 2.

1-Varianza/Tiempo
2-Rango Reescalado
3- Diagarma Log-Escale
4- Varianza/Octavas

Resaltados en cursiva, se encuentran los resultados de la estimación de *H*, por el método del diagrama log-escala implementado mediante la script LDestimate de Darryl Veitch, que es el de mayor precisión y menor sesgo.

Referencias Bibliográficas

[1]http://postgrado.info.unlp.edu.ar/Carreras/Magisters/Redes_de_Datos /Tesis/Scappini_Reinaldo.pdf
[2] http://www.cubinlab.ee.unimelb.edu.au/~darryl/secondorder_code.html
[3] [Kihong Park, Walter Willinger Self-Similar Network Traffic and Performance Evaluation Copyright 2000 John Wiley & Sons, Inc. ISBNs: 0-471-31974-0 (Hardback); 0-471-20644-X (Electronic).
[4][M. Alzate y A. Monroy, Uso de la transformada wavelet para el estudio de tráfico fractal en redes de comunicaciones. Revista Ingeniería Vol. 7 No. 1 Junio 2002 Páginas: 11 –24. Universidad Distrital Francisco José de Caldas.
[5] [W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of ethernet traffic (extended version), IEEE/ACM Transactions on Networking, vol.2, pp.1–15, Feb. 1994.
[6] http://imdc.datcat.org/collection/1-055M-0=WIDE-TRANSIT+100+Megabit+Ethernet+Trace+2007-01-09+(Anonymized)
[7] http://pma.nlanr.net/Special/sdsc1.html
[8] http://pma.nlanr.net/Special/
[9] http://www.caida.org/home/
[10] http://www.wireshark.org/

IP Core Para Redes de Petri con Tiempo

Ornaldo Micolini[1], Julián Nonino[1] y Carlos Renzo Pisetta[1]

[1]Facultad de Ciencias Exactas Físicas y Naturales
Universidad Nacional de Córdoba, Argentina
omicolini@compuar.com, {noninojulian, renzopisetta}@gmail.com

Abstract. En este trabajo, se presenta un procesador de Redes de Petri con Tiempo, el que es la evolución del Procesador de Petri Temporizado. Este procesador es programado directamente con las matrices y vectores del formalismo de Petri, lo que permite aprovechar el poder de las redes de Petri para modelar sistemas de tiempo real y verificar formalmente sus propiedades, evitando errores de programación al implementar el programa a ejecutar.
Este desarrollo ha sido realizado como un IP-cores y es usado en un sistema Multi-core. De esta manera, es posible realizar la implementación del sistema utilizando este IP-core, lo que asegura las propiedades del modelo realizado con la red de Petri con Tiempo, que verifican los requerimientos del modelo que representa al sistema real, sean cumplido.

Key words: Multi-core, Red de Petri, Procesador

1 Introducción

Los sistemas informáticos son complejos tanto en su estructura como en su comportamiento, más aun cuando tienen un gran número de estados y numerosas combinaciones de datos y eventos de entrada.

Abordar soluciones de sistemas complejos y crítico, para dar solución a sistemas en tiempo real, tiene problemas como: la complejidad inherente de la especificación, la coordinación de tareas concurrentes, la falta de algoritmos portables, entornos estandarizados, software y herramientas de desarrollo.

Y teniendo en cuenta, las tendencias inequívocas en el diseño de hardware, que indican que un solo procesador no puede ser capaz de mantener el ritmo de incrementos de rendimiento. Por lo que la evolución de los procesadores, que es consecuencia de la mayor integración y la composición de distintos tipos de funcionalidades integradas en un único procesador. Más aun, hoy la disponibilidad de transistores ha hecho factible construir en una sola pastilla varios núcleos de procesador que ha resultado en el desarrollo de la tecnología Multi-core [1].

La obtención de rendimiento decreciente del paralelismo a nivel de instrucción (ILP) y el costo del incremento en la frecuencia debido principalmente a las limitaciones de potencia (se sugiere que un 1% de aumento de velocidad de reloj resultados en un aumento de potencia del 3% [2]) ha motivado el uso de los Multi-core.

Por lo cual los procesadores Multi-core son una propuesta para obtener aumento de rendimiento. Lo que se traduce principalmente en menores tiempos de ejecución, consumo ruido, densidad de energía, latencia y más ancho de banda en las comunicaciones inter-core. Si también consideramos a los Multi-core heterogéneos que tienen como ventaja emplear cores especializados, diseñados para tareas específicas. Es decir, optimizado según la necesidad. Estos tienen la capacidad de usar los recursos de hardware disponibles donde el software específicamente lo requiere. [3]

Con el fin de aumentar el desempeño, estos sistemas hacen uso colaborativos de multi-hilos y/o multi-tarea, lo que permite aprovechar los múlti-núcleos. Pero se requiere de más trabajo en el diseño de las aplicaciones, ya que emergen con fuerza la problemática de los sistemas concurrentes.

Por lo que con estos procesadores, la programación paralela es indispensable para la mejora del desempeño del software en todos los segmentos de desarrollo y con más razón en el segmento de sistemas de tiempo real.

Para dar solución a los sistemas reactivos, paralelos y de tiempo real, en relación con los siguientes aspectos:

- Problemas de concurrencia que emergen en la programación paralela, por no ser componible, es decir, no se puede obtener un programa paralelo de la composición directa de dos programas secuenciales.
- Que el hardware de soporte a la implementación de sistemas concurrentes, permitiendo mejorar los algoritmos paralelos.
- Asegurar los requerimientos temporales en los sistemas de tiempo real, es decir, los intervalos mínimos y máximos para la ocurrencia de un evento. Para lo cual el hardware facilite la programación de estas restricciones en forma directa.
- Tareas de codificación, que se requieren para la implementación de un modelo, conducen a errores e incrementan el esfuerzo, por lo que es muy valorable que no exista ninguna tarea entre el modelo y el software a ejecutar.

2 Objetivo

2.1 Objetivo Principal

El objetivo principal de este trabajo es diseñar e implementar un procesador de Redes de Petri con Tiempo, que ejecute la semántica temporal y se programe en forma directa a partir de las ecuaciones de estado del modelo.

2.2 Objetivos Secundarios

Los objetivos secundarios de este trabajo son:

- Describir brevemente las Redes de Petri con Tiempo con el fin de realizar su implementación por hardware.
- Mantener la ejecución de las Redes de Petri ordinarias con parámetros temporales en dos ciclos de reloj.
- Implementar el procesador de Redes de Petri en un IP-core.

3 Redes de Petri con Tiempo

En estas redes, cada transición con tiempo tiene asociado un intervalo de tiempo [a, b] que establece el intervalo de tiempo dentro del cual puede ser disparada la transición, con el fin de homogenizar las definición matemática definimos transiciones inmediatas con límite inferior cero. [4]

3.1 Definición Matemática

Una Red de Petri con Tiempo (TPN) [5] y marcada, se define matemáticamente como una 8-tupla de la siguiente manera:

$$\{P, T, I^+, I^-, H, C, m_0, IS\}$$

Donde $\{P, T, I^+, I^-, H, C, m_0\}$ es una red de Petri plaza transición marcada con brazos inhibidores y plazas acotadas, y IS es la función estática de intervalos $[a, b]$ asociados a cada transición.

Dónde:

P: es un conjunto finito y no vacío de plazas.

T: es un conjunto finito y no vacío de transiciones, P y T son conjuntos disjuntos

$\boldsymbol{I^+}$, $\boldsymbol{I^-}$: son las matrices de incidencia positiva y negativa. La matriz ***I*** es las diferencias entre $\boldsymbol{I^+}$, $\boldsymbol{I^-}$.

$$PxT \rightarrow Z$$

H: es la matriz de brazos inhibidores.

$$PxT \rightarrow \{0,1\}$$

C: es el vector de cota de plaza

$$C \rightarrow N$$

IS: es la función estática de intervalos asociados a cada transición.

$$T \rightarrow \mathbb{Q}^+ \times (\mathbb{Q}^+ \cup \infty)$$

La función IS asocia a cada transición un par de valores que representan los límites temporales máximo´ y mínimo entre los cuales la transición podrá ser disparada. De manera tal que

$$IS(t) = [min, max] \; \forall \; t \in T$$

Como la función IS representa un intervalo temporal, para cada transición ***t*** sensibilizada se introduces el valor $timer_t$, que se auto incrementa con el tiempo, si la transición esta sensibilizada y se cumplir: $\min \leq timer_t \leq max$ el disparo sea posible.

Estas cotas deben cumplir las siguientes condiciones:

- $0 \leq \min < \infty$
- $0 \leq \max \leq \infty$
- $min \leq \max si \max \neq \infty$
- $min < \max si \max = \infty$

Al valor min lo lamamos Earliest Firing Time EFT (Instante de disparo más temprano). Y, al valor max se le llama Latest Firing Time LFT (Instante de disparo más tardío).

Existen dos tipos de intervalos destacables:

- Intervalo puntual $[a, a]$. En este caso, el tiempo de disparo es fijo, después de sensibilización se espera un tiempo a.
  Un disparo inmediato es representado por α = 0 y se comporta como en las Redes de Petri plaza transición.
- Intervalo sin restricción temporal, $[a, \infty]$. Se disparara en algún momento después de sensibilizarse y un tiempo a.

3.2 Estados en una Red de Petri Temporizada

En las Redes de Petri con tiempo, el estado de la red es definido por el vector de marcado m_i y por el vector de valores de intervalos de transición $timer$ de la red, que lleva la cuenta de tiempo de cada transición sensibilizada. Por lo tanto el estado es:

$$E = (mi, timer)$$

3.3 Transición Sensibilizada y Disparo de una Transición

Cundo nos referimos a una transición hay que distinguir las siguientes cuestiones: transición habilitada o sensibilizada, transición no habilitada y disparo de una transición.

En una Red de Petri marcada, con una marca $\boldsymbol{m_k}$, se dice que una transición t_j se encuentra habilitada o sensibilizada si y solo si (sii) todos los lugares del conjunto de plazas $\bullet\, \boldsymbol{t_j}$ de entrada a la transición tienen al menos la cantidad de marcas igual al peso de los arcos ($\boldsymbol{w(p_i, t_j)}$) de entrada a la transición $\boldsymbol{t_j}$, esto es:

$$p_i \in \bullet\, t_i, m(p_j) \geq w(p_i, t_j)$$

Si el $timer$ de la transición es cero, se debe habilitar $timer_t$ para que se auto incremente con el tiempo.

Las transiciones sensibilizadas pueden ser disparadas en el intervalo [a, b], y su disparo provoca un nuevo marcado es decir un cambio de estado. La ecuación para calcular el cambio de estado o la nueva marca alcanzada por el disparo de $\boldsymbol{t_j}$ es $\partial\,(m_k, t_j)$, y se define por la siguiente expresión:

$$\partial\,(m_k, timer_t) = \begin{cases} m_{k+1}(p_i) = m_k(p_i) - w_{ij} & ,\forall\, p_i \in t_j \bullet \\ m_{k+1}(p_i) = m_k(p_i) + w_{ji} & ,\forall\, p_i \in t_j \bullet \\ \min \leq timer_t \leq max & \\ m_{k+1}(p_i) = m_k(p_i) & ,en\ el\ resto\ de\ los\ casos; \end{cases}$$

$$min = a, max = \mathrm{b}$$

Donde el $timer_{t_j}$ se incrementa en cada ciclo de reloj mientras la transición se encuentra sensibilizada.

4 Arquitectura y Funcionamiento del Procesador de Petri con Tiempo

El procesador ejecuta la ecuación de cambio de estado resolviendo solo un disparo de una transición a la vez, esto permite resolver todos los casos de disparos, los simples (disparo único) y los disparos múltiples, realizándolos como una secuencia de disparos simples, esto simplifica el hardware.

Las disparos son transmitidos por los hilos que se ejecutan en los cores a través del bus del sistema, según las solicitudes emergentes del sistema que se está ejecutando. Esto disparos son recibidos por el Procesador de Petri con Tiempo y almacenado en la cola de disparos de entrada. Existe una cola FIFO por cada transición, la salida de este conjunto de colas es una palabra con tamaño igual a la cantidad de transiciones, la cual tiene unos en las posiciones correspondientes a las transiciones con disparos solicitados, el orden del bit en la palabra es igual al número de la transición que solicita el disparo. Los bits, que se corresponden con las transiciones que no tiene disparo solicitado, son cero, es decir no hay solicitud de disparo.

La cola de salida tiene una estructura similar, pero comunica los disparos resueltos a los hilos.

En la Fig. 1 se muestran los distintos módulos que componen el procesador, resaltando las principales diferencias con versiones anteriores. [6]

El módulo de I/O Datos gestiona el acceso de los cores a las matrices y vectores que programan el sistema.

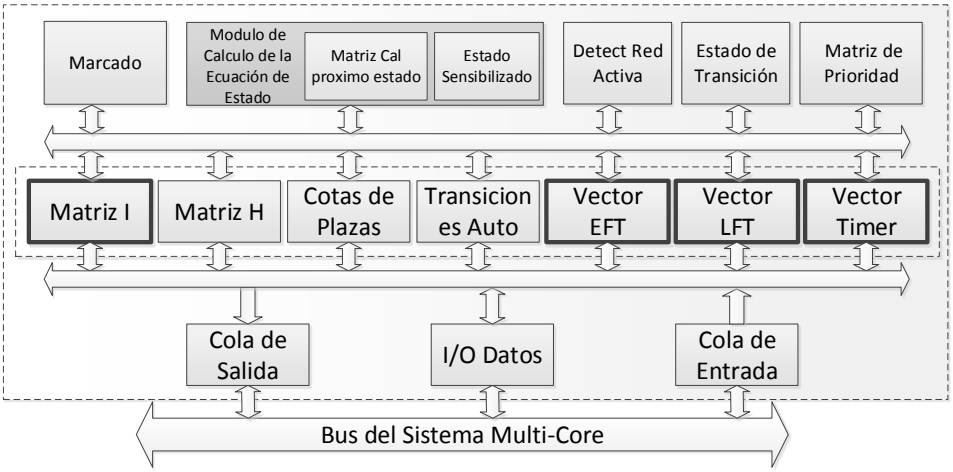


Fig. 1. Procesador de Petri con Tiempo.

El programa del sistema son las matrices y vectores descriptas en la ecuación de estado, esto permite programar el procesador en forma directa a partir de la Red de Petri con Tiempo.

Aquí se han agregado la matriz de Brazos Inhibidores y el vector de Cota de Plaza que no figuran en la ecuación de estado presentada en este trabajo, pero son los mismos que en el Procesador de Petri presentado en otros trabajos [7].

La responsabilidad del Módulo de Cálculo de la Ecuación de estado es la siguiente:

1. Calcular el nuevo estado que resultaría por disparar solamente una transiciones una vez, por lo que resultan tantos vectores de estados calculados como transiciones, y se almacenan. Esto se realiza en paralelo sumando al estado actual a cada columna de I y almacenando todos los vectores resultantes, los que serán evaluados para determinar si son los posibles nuevos estado. Esta operación es realizada siempre que cambia el estado del procesador, vector Marcado.
2. Determinar que transición esta sensibilizada. Se toma todos los vectores calculados en 1 y se verifica que se cumpla que ninguna plaza tenga marcado negativo y tampoco supere la cota de plaza, estas son las transiciones sensibilizadas.
3. Se arranca o para los Timer_t . Si en una transición sensibilizada $\text{Timer}_t = 0$ se arranca Timer_t y si $\text{Timer}_t \neq 0$ no se hace nada.
4. Disparo de una transición. Las transiciones que cumplen con:

$$Vector\ EFT \ \leq \ \text{Vector Timer}_t \ \leq Vector\ LFT$$

   Las transiciones que cumplen con esta condición y han recibido por la cola de entrada un disparo o el disparo están programado como automático, conforman un conjunto de disparos posibles

   De este conjunto se selecciona el de mayor prioridad y se ejecuta la transición.

   Según la transición ejecutada se actualiza el vector de estado, y se pone Timer_t a cero.

5. Se ejecuta como un ciclo continuo los pazos 1, 2, 3 y 4.

El sistema posee una unidad que detecta cuando ninguna transición esta sensibilizada y Vector Timer supera el tiempo máximo; esta condición genera una interrupción que comunica que el sistema ha finalizado o esta interbloqueado, esta característica es de suma utilidad para verificar el diseño e implementación del sistema.

La Tabla 1 muestra las diferencias significativas, desde el punto de vista de la ejecución de las distintas semánticas, estas son:

Tabla 1. Comparación entre Semánticas Temporales.

		Con Tiempo	Temporizada
1	Interrumpible	Si	No
2	Representa las dos semánticas	Si	No
3	Matrices usadas	I	I+, I-
4	Permite contener subredes	No	Si

De este cuadro se desprenden las siguientes observaciones:

1. Siendo que las TPN son interrumpibles y las Redes de Petri Temporizadas (TdPN) no lo son, para el caso de múltiples disparos y transiciones en conflicto, un TPN lo resuelve según el intervalo de tiempo; en cambio una TdPN lo hace explícitamente en la matriz de prioridad. Esto hace más complejo el modelado con TdPN e indispensable incluir en el procesador una matriz de prioridades.

Dado que la mecánica de ejecución de las TdPN requiere de un estado más para no ser interrumpibles los tokens son retirados inmediatamente de la plaza y no pueden ser solicitados por otra transición.

2. Dada una red con TPN, una transición, que por semántica es interrumpible, puede transformarse en una no interrumpible modificando la red. Esto se logra encerrando con una transición inmediata la transición temporiza. Lo que tiene como impacto un incremente de una plaza y una transición adicional por cada transición no interrumpible.
3. Para realizar el cálculo de un nuevo estado las TPN lo hacen con una matriz de enteros con signo mientras que las TdPN lo realizan con dos matrices de enteros sin signo; por lo cual debemos analizar dos casos:
    a. Si los pesos de los arcos son uno:
        i. Las TPN requieren de una matriz con 2 bit por elemento.
        ii. Las TdPN requieren de dos matrices binarias.
    b. Si los pesos de los arcos son uno o mayor a uno:
        i. Las TPN requieren de una matriz de enteros con signo.
        ii. Las TdPN requieren dos matrices de enteros sin signo.

    En el primer caso los recursos utilizados son similares. Por lo que la selección de uno u otro procesador depende de la semántica a utilizar. Mientras que, en el segundo caso los recursos utilizados por las TdPN son mayores. La ventaja de una con respecto a la otra en cuestión de recursos está determinada por incremento de la matriz de incidencia dada por la conversión de las transiciones Time a su equivalente no interrumpibles.
4. El procesador que implementa la semántica TdPN utiliza dos estados para realizar el cálculo de los toquen que entran de una transición y los que salen de esta. Esta diferenciación de estados nos permite insertar una nueva red de Petri entre los dos estados de una transición, lo que posibilita que el procesador puede ser extendido a redes de Petri jerárquicas; ya sea haciendo uso de la semántica TdPN o de las redes de Petri ordinarias. Esto en la actualidad es motivo de una nueva investigación.

Las dos semánticas son investigadas, puesto que las TPN requieren de menos recursos para resolver problemas no interrumpibles (que son los más habituales). Mientras que las TdPN presentan potencial de mejora al permitir construir redes de Petri jerárquicas. [8].

5 Análisis de Rendimiento

La implementación de sistema ha sido realizada en una plataforma Atlys™ Spartan-6, los cores utilizados son los MicroBlaze ver8.40 [9] que ejecuta un Sistema Operativo XilKernel ver5.01a. Interconectado con el Procesador de Petri Temporizado por un bus AXI [10].

Para comprobar correcto funcionamiento del IP Core y analizar los tiempos de sincronización, se realizaron mediciones para distinto número de iteraciones y numero

de hilos tratando de acceder a una variable compartida en exclusión mutua. Luego se compararon el Procesador de Petri con una implementación utilizando semáforos, ambos resolviendo un mismo problema. La elección de este segundo método de sincronización se basa en que son el mecanismo más ligero para realizar éstas tareas.

A partir de estas mediciones se calculó el Speedup, los resultados se muestran en la Fig. 2, se puede observar que, para todos los casos, el procesador de Petri es en promedio es entre un 15% y un 30% más rápido que el uso de semáforos para resolver el problema de sincronizar múltiples hilos que desean escribir sobre una variable compartida e incluso, se alcanzan picos de hasta un 70%.

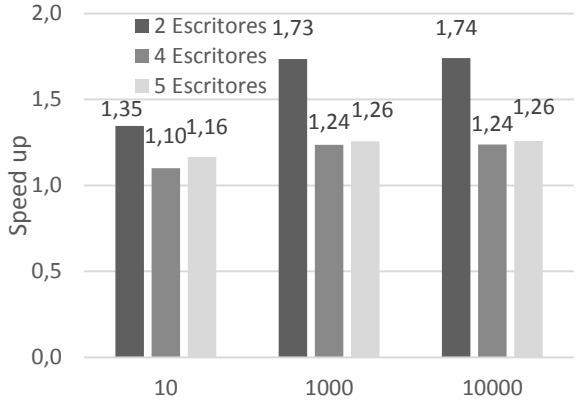


Fig. 2. Tiempos de sincronización por iteración

Estas mediciones se realizaron con tiempos $EFT\ y\ LFT$ cero, de manera que el rendimiento es el mismo obtenido en el procesador de Redes de Petri sin la semántica temporal. Esto es válido ya que el tiempo de una transición es parte del modelo, es decir, es el mismo para el procesador de Petri como para la implementación con semáforos y el propósito es medir únicamente los tiempos de sincronización.

Además, como se observa en la Fig. 3, el procesador necesita únicamente un semi-ciclo de reloj, desde que el contador alcanza el valor EFT hasta que el disparo se coloca en la cola de salida. La demora introducida es despreciable en relación con el tiempo que tiene un δt de un ciclo de reloj.

Teniendo en cuenta lo despreciable de la latencia y tomando el tiempo como parte del modelo es posible analizar el rendimiento sin tener en cuenta los vectores EFT y LFT.

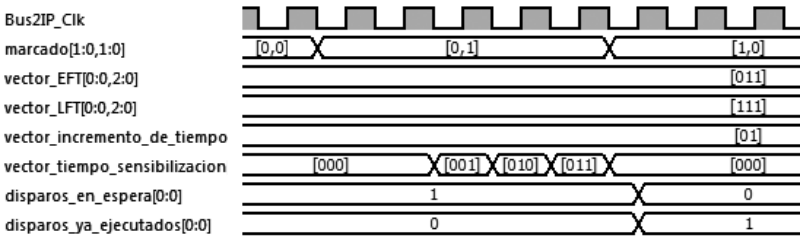


Fig. 3. Ejecución en hardware.

6 Crecimiento del IP Core

Se analizó el crecimiento del procesador en función de los parámetros que posee. Para esto se generaron procesadores de 8x8, 16x16, 32x32, 48x48 y 64x64 (Plazas por Transición) con capacidad de 7 bits por plaza y elementos de tiempo de 48 bits y se graficaron los resultados, los que se pueden observar en la Fig. 4.

Se observa que el crecimiento del IP Core no es algo para despreciar, puesto que la cantidad de elementos empleados crese rápidamente con el producto de las Plazas por las Transiciones.

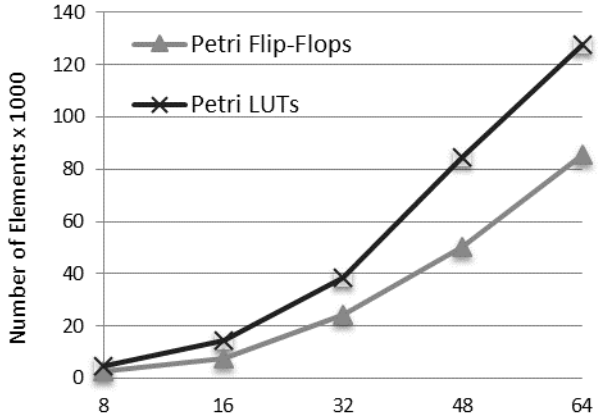


Fig. 4. Crecimiento del IP Core

Por otra parte, ya que es posible sintetizar un procesador para cada semántica es deseable determinar y comparar el consumo de recursos para cada uno. La Fig. 5 muestra la comparación del crecimiento entre las distintas implementaciones.

Se puede observar que ambos procesadores utilizan aproximadamente la misma cantidad de Flip-Flops pero la implementación para redes temporizadas utiliza un 90% mas LUTs para el mismo número de plazas y transiciones.

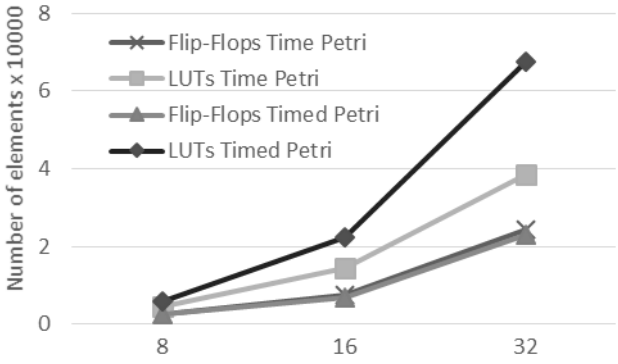


Fig. 5. Recursos usados por distintas semánticas.

7 Conclusión y Aportes

En el presente trabajo, se desarrolla el Procesador de Petri con Tiempo, que permite desacoplar los la concurrencia del procesamiento secuencial. Teniendo en cuanta que el Procesador de Petri con Tiempo permite utilizar Redes de Petri Temporizadas, este procesador puede remplazar a su predecesor y preserva sus particularidades.

El modelo de Petri es adecuado para implementar, validar y verificar un sistema paralelo con concurrencia, este tiene una representación algebraica que este procesador usa como el código ejecutable. Las ventajas de este procesador son la disminución de:

- Esfuerzo de programación, la ecuación de estado es ejecutada directamente en el procesador, y no se requiere programación adicional.
- El gap entre las restricciones temporales y sus programaciones. Puesto que se trata de los vectores temporales propios de la semántica usada por el procesador.

Referencias

1. Hennessy, John L. Computer Architecture A Quantitative Approach.: Denise E. M. Penrose, 2007.
2. Domeika, M. Software Development for Embedded Multi-core Systems. 0 Corporate Drive, Suite 400, Burlington, MA 01803, USA : Linacre House, Jordan Hill, Oxford , UK., 2008.
3. Sundararajan Sriram, S. S. B. EMBEDDED MULTIPROCESSORS, Scheduling and Synchronization. Boca Raton, 2009.
4. Ramachandani, C. Analysis of Asynchronous Concurrent Systems by Timed Petri Nets. Cambribge, Massachussets : Massachussets Institute of Technology, 1974.
5. Izquierdo, García. Modelado e implementación de sistemas de tiempo real mediante redes de petri con tiempo. Zaragoza, 1999.
6. IP Core for Timed Petri Nets. Micolini, Orlando, Nonino, Nulián and Pisetta, Carlos R. Buenos Aires, Argentina: s.n., 2013. CASE 2013,unpublished
7. Procesador de Petri para la Sincronización de Sistemas Multi-Core Homogéneos. Micolini, Orlando, y otros. Buenos Aires, Argentina : s.n., 2012. CASE 2012.
8. Jensen , Kurt y Kristensen, Lars M. Coloured Petri Nets: Modelling and Validation of Concurrent Systems. New York : Springer, 2009 .
9. Xilinx. MicroBlaze (UG708). 2012.
10. AXI Interconnect (DS768). 2012.

Analysis of Radio Communication Solutions in Small and Isolated Communities under the IEEE 802.22 Standard

A. Arroyo Arzubi[1], A. Castro Lechtaler[1 y 3], A. Foti[4], R. Fusario[1, y 4], J. García Guibout[2] and L. Sens[4]

[1] Escuela Superior Técnica - Facultad de Ingeniería del Ejército - IESE, C1426, Buenos Aires; [2]Instituto Tecnológico Universitario - Universidad Nacional de Cuyo, M5500, Mendoza; [3]Universidad Nacional de Chilecito, F5360, Chilecito, La Rioja; [4]Universidad Tecnológica Nacional, C1042, Buenos Aires; República Argentina.

{A. Arroyo Arzubi, arroyoarzubi@iese.edu.ar, A. Castro Lechtaler, acastro@iese.edu.ar, A. Foti, foti.antonio@gmail.com, R. Fusario, rfusario@speedy.com.ar, J. García Guibout, jgarcia@itu.uncu.edu.ar, L. Sens, lsens@frba.utn.edu.ar}

Abstract. In recent years the use of wireless communications has increased significantly. Rural communities without cable network communication have found a solution in wireless technologies. Based on previous fieldwork, this paper analyzes software development of integration based technologies for communication equipment. It focuses on the feasibility of the IEEE 802.22 standard as a solution to the wireless problem.

Keywords: IEEE 802.22, White Spaces, Cognitive Radio, Rural Communications, Digital TV Broadcast.

1 Introduction

In the framework of the Project ***Communitarian Private Networks*** [1], different technologies providing links to small and isolated communities have been analyzed and compared. These communities, with low population densities, hold no commercial interest to service providers [2], [3], [4]. Notwithstanding, several rural facilities maintain operations in these isolated areas, providing significant quantities of food products at different stages of manufacturing. They supply not only nearby cities, but also constitute an important source of export commodities and revenue for many countries.

The geographic dispersion of these facilities interfere with cable communications – either with copper pairs, coaxial or optic fiber cables – due to high costs and maintenance problems. Consequently, the solution consists of establishing full duplex links via radio waves at a 30 to 70 km distance between antennas and at frequencies not restricted by government regulations [5], [6].

Towards the end of the 90s and beginning of this century, technical problems evolved side by side with their solutions. The process lead to the approval, on July 1st, 2011, of the standard IEEE ***802.22 - Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Policies and Procedures for Operation in the TV Bands*** [7].

The present work analyzes the use of this type of links in the same area where our group is conducting fieldwork.

2 Previous Fieldwork and Testing of New Technologies

The Project *Communitarian Private Networks* [5] explores different technologies providing communications to small and isolated rural communities with low population densities and without telephone services, whether these may be landlines or cellular. Hence, these communities do not have access to voice, data or internet networks.

A small community which met the requirements of the Project was searched: an isolated and distant town where experiences can be appropriately implemented. The community Corral de Lorca, in the department of General Alvear, province of Mendoza was finally selected. Its location is shown in Figure 1.

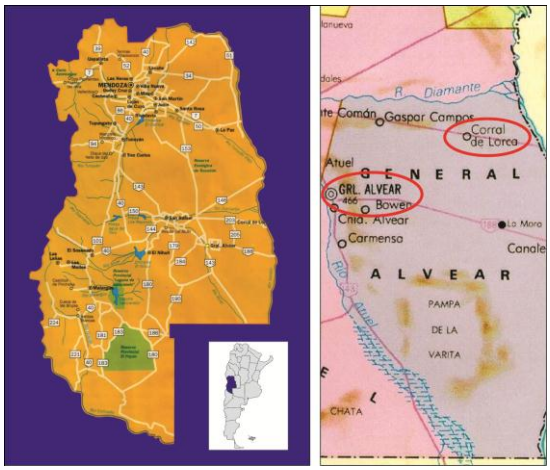


Figure 1. Geographic Location of Corral de Lorca

The technologies under study were: Power Line Communications (PLC) [2, 3] and the standard 802.11 [4]. Both experiences show that PLC technology is not recommended for outdoor links under the required conditions.

The considered solution involves establishing a point to point link where one endpoint is located in the department´s main city, General Alvear, with access to the PSTN network, and the other in the Corral de Lorca community, located in the southwest of the province of Mendoza, Argentina.

Two Motorola Canopy platforms are used at bandwidths of 2.4 GHz and 5.7 GHz (similar to WiMax). At the Corral de Lorca node [6], phone services can be implemented through VoIP as well as 802.3 to the Local Area Network and 802.11 for Wireless Internet Services.

The link is placed at a critical distance. Corral de Lorca is 70 kilometers from General Alvear in a straight line over a desert but with dense vegetation close to the base station area. To analyze the propagation issues the freeware ***Radio Mobile*** developed by Roger Coude [8] is used.

The outcome is not entirely satisfactory. Significant attenuation is registered due to the distance of the trans-horizon link and to the strong attenuation resulting from the dense vegetation at the outskirts of General Alvear. These factors contribute to a low value in the signal/noise ratio at the reception end in Corral de Lorca.

The conclusions are:

- The link is studied as a typical case to solve the general problem of rural populations. Hence, it could be generalized later for the application of analogous situations. In these cases, the distance to cover should range between 60 to 70 kilometers.
- Coordinates and terrain conditions are detailed between the two endpoints. Estimates for different bandwidths are established, i.e. 2.4 GHz and 5.7 GHz.
- The distance between the endpoints (60 km) is greater than regular distances contemplated in the theory for the application of standard 802.11.
- In addition, the analysis focuses on the fact that, in practice, transmitted signals in rural areas behave differently from those in urban settings because the former suffer less from noise spectrums. Radio Mobile software is considered to be a valuable tool in the design of radio electric links.

As a result of these experiences, continuing work is focusing on the new 802.22 standard.

3 Technical Problems and New Solutions

3.1 Introduction

The boundaries between the fields of communications and computer science have merged over time. Several concepts used in the field of telecommunications are now encountered in computer science and vice versa. Also, methodological practices of computer science are an integral part of telecommunications nowadays.

Consequently, today we refer to both disciplines as Information and Communications Technology - ICT[1] or as Teleinformatics, as referred to by European and American scholars.

Moreover, by the end of the 90s and beginning of this century, wireless communications increased exponentially. For instance, currently the total number of mobile phone users exceeds the number of existing landlines.

The pervasive use of mobile communications presents several technical difficulties, which in turn lead to the development of their consequent technical solutions. In the following section, the main changes of the advance in wireless technology are outlined.

[1] Also known as TICs in Spanish

3.2 White Space and Congestion Bands

Modern societies are increasingly relying on radio spectrum use. The pervasiveness of wireless services and communication devices (mobile phones, police communications, Wi-Fi and digital TV broadcasting) are examples of this dependency. It has become one of the most necessary resources of modern times [9].

Global demand growth for mobile data traffic has increased between 2011 and 2012 at a rate over 100%. The expected growth rate of this demand for the period 2008 and 2013 is estimated to average at 131% per year [10], exceeding 2.000.000 Terabytes per month by the end of the current year. The intense spectrum use, up to 5 GHz, and more specifically at the coverage below 1 GHz, has lead to a thorough review of regulatory policies, along with a renewed interest in ***White Space*** research [2] [11].

Possible solutions to the increasing traffic, especially below 1 GHz, are: review and redesign of the regulatory framework, reduction of wireless services broadcasting, improved compression standards, replacement of various services by satellite or cable, dynamic spectrum access, and development of cognitive radio technologies. The latter is oriented to take advantage of under-utilized frequencies, temporary voids of primary signals, and different types of white space.

The CEPT, ***European Conference of Postal and Telecommunications Administrations***, has defined ***White Space*** as "a label indicating a part of the spectrum, which is available for radio communication applications (service or system) at a given time in a given geographical area on a non-interfering or non-protected basis with regard to other services with a higher priority on a national basis" [12]. Currently, several research efforts from different organizations, national and international, are working on white space.

Cognitive Radio Technology (CRT) is considered another possibility to address the rising spectrum shortage. When fully operational, CRT could provide technologies for a variety of applications (rural broadband, public safety and emergency response, and urban frequency use). This technology will also have significant consequences for dynamic detection and spectrum management.

3.3 Software Defined Radio

With the exponential growth of the ways and means by which people need to communicate through wireless communications, modifying radio devices easily and cost-effectively has become critical.

The technology ***Software Defined Radio (SDR[3])*** provides flexibility and profitability, as well as grants end users with comprehensive benefits from service providers and product developers *[13]. The* ***Wireless Innovation Forum*** defines ***Software Defined Radio*** as "radio in which some or all of the physical layer functions are software defined."

The radio is a device which transmits or receives wireless signals using a portion of the radio spectrum. Traditional radio devices exclusively based on hardware (e. g.:

[2] Or white holes.
[3] Also known as **S*oftware Radio***.

mixers, filters, amplifiers, modulators/demodulators, and detectors) are limited because their features can be modified only by physical intervention.

On the other hand, a ***Software Defined Radio (SDR)*** is implemented by means of software on a computer or embedded system. The concept is not new, but the rapidly evolving capabilities of digital electronics render practical many processes which used to be only theoretically feasible before [13].

Under this technology, the software proves to be efficient at a relatively inexpensive cost, with multimode and multiband wireless devices which can be continuously improved with software updates. In some cases, the software manages some or all of the functions to operate the radio equipment (including those of the physical layer processing).

3.4 Cognitive Radios [4]

At the end of the decade of the 90s, Joseph Mitola[5] and Gerald Maguire, researchers from the Royal Institute of Technology[6] developed what they called ***Cognitive Radio,*** an improvement of their previous work on ***Software Defined Radio*** technology [14] [15],

While Software Defined Radio offers great potential, it also requires arduous processing, limiting its flexibility and adequacy of network response.

Cognitive Radio embedded in communications software, such as ***Radio Knowledge Representation Language - RKRL***, can be considered an intelligent and efficient system for radio communications and protocols. Basically, it provides mechanisms based on the use of smart technology to optimize the spectrum.

As mentioned in 3.2, the allocation of frequencies in a saturated spectrum is not optimal, originating ***White Space***. A special range is assigned to the operators for the use of ***Digital TV Broadcasting***.

Those were the reasons which led to develop Cognitive Radio for wireless communications: ***to detect the parts of the radio frequency spectrum used inefficiently and to allow reuse without causing interference with the services assigned to them.*** The solution of these problems by variable frequency allocation, allows others to take advantage of unused parts of the spectrum.

Using intelligent software, Cognitive Radio periodically scans the spectrum in search of white holes, detects the use given to each of them, and then determines whether it is reusable.

The system operates by changing the transmitter parameters based on interaction with the environment. It has the ability and the technology to capture or sense the information from other radio equipment, providing spectrum awareness whereas reconfigurability enables the radio to be dynamically programmed.

It can be programmed to transmit and receive on a variety of frequencies and to use different transmission access technologies supported by its hardware design.

---

[4] Mitola defines ***cognitive*** as ***the mix of declarative and procedural knowledge in a self-aware learning system***.

[5] Joseph Mitola III received his doctorate in the Royal Institute with his thesis ***Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio***.

[6] Located in Stockholm, Sweden.

These operating procedures show the interaction between hardware design and application software development. They also represent a typical teleinformatics application, as characterized by Minola in his thesis.

3.5 Digital TV Broadcasting

Frequency spectrum use for TV broadcasting has varied since the first black and white broadcasts to the current digital high definition systems. Two bands are used: VHF (54 to 88 and 174 to 216 MHz) and UHF (512 to 806 MHz).

Figure 2a. Province of Buenos Aires

Figure 2b. Rest of the country

In several South American countries, the Japanese standard **Integrated Services Digital Broadcasting** (**ISDB**) has been adopted with a few variants, such as the replacement of the compressing system MPEG-2 for MPEG-4[7]. It was developed by the ***Association of Radio Industries and Businesses,*** known as ***ARIB,*** *w*hich promotes the efficient use of the spectrum.

ISDB include four standards depending on the used medium: ISDB-S (satellite), ISDB-T (terrestrial), ISDB-C (cable) and 2.6 GHz mobile broadcasting. All of them are based on multiplexing with a transport stream structure and are capable of High-Definition Television (HDTV) and standard definition television. The name of the standard was chosen for its similarity to ISDN (Integrated Services Digital Network). Both allow the simultaneous transmission of multiple channels of data through the multiplexing method.

In most cases, broadcasting stations have antennas reaching about 150 meters high, with significant coverage areas.

In the case of Argentina, more that 50 broadcasting stations have been set up as of July of 2013, covering a significant area of the country. The plan is to cover practically all of the populated areas, giving service to 90% of the population. Figures 2a and 2b illustrate the cities where these stations have been installed.

[7] ***International Services Digital Broadcast***, Terrestrial Brazilian version ISDB-Tb.

3.6 IEEE 802.22 as a Solution for Rural Areas

In Argentina, as in many countries with large rural areas, most cities are located within a range of 40 to 80 km apart in average.

The Project *Communitarian Private Networks* focuses on evaluating solutions to the communication problems of rural areas, in particular, isolated communities with low population density.

In our countries, the intensive use of spectrum and saturation in many of its frequency bands is due to wireless communications which has been the only feasible solution.

The 802.22 standard aims at using the vacancies in the TV spectrum. These frequencies are particularly suitable for remote areas where cables signal transmission are expensive or difficult to implement. Cables could only be replaced by costly satellite services. Thus, to implement a link using spare frequencies in these bands may be a practical and inexpensive solution.

In our country, the TV on the VHF band will be eliminated in 2016 (analogic blackout), liberating most of the UHF band, considering that the ***Argentine National Authority for Broadcasting Services - AFSCA***[8] has licensed only a few channels in the main cities (22 to 36).

As the new digital TV technology allows several standard definition programs in the same bandwidth of one high definition channel, there is a significant spectrum saving, and we still can get lots of free frequencies (channels 38 to 69), mainly in small cities.

It is an opportunity for this IEEE 802.22 standard to be considered in the spectrum reallocation under study by the ***National Argentine Spectrum Authority - CNC***[9].

4. IEEE 802.22. Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY)

4.1. Introduction

On July 1st 2011, the standard ***IEEE 802.22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Policies and Procedures for Operation in the TV Band*** was approved under the sponsorship of the LAN / MAN Standards Committee.

The standard aims to set up criteria for the deployment of multiple interoperable products of the 802.xx series[10], offering fixed broadband access in various geographical areas, including especially those of low population density in rural areas, and avoiding interference to services working in the television broadcasting bands.

The standard, commonly known as ***Wireless Regional Area Networks – WRANs***, has been developed to operate primarily as broadband access to data networks in rural areas.

[8] ***Autoridad Federal de Servicios de Comunicación Audiovisual***.
[9] ***Comisión Nacional de Comunicaciones***.
[10] Wireless.

4.2. General features

The standard includes cognitive radio techniques to moderate interference to other existing operators, to grant geolocation capability, to provide access to a database of incumbent services, and to detect the presence of other services through spectrum-sensing technology, such as different WRAN systems or IEEE 802.22.1[11] wireless beacons.

The WRAN systems involve the use of channels ranging from 54 to 862 MHz in the VHF and UHF bands. The use of cognitive radio technologies scans for spare frequencies while avoiding interference with TV stations operating in the same bands.

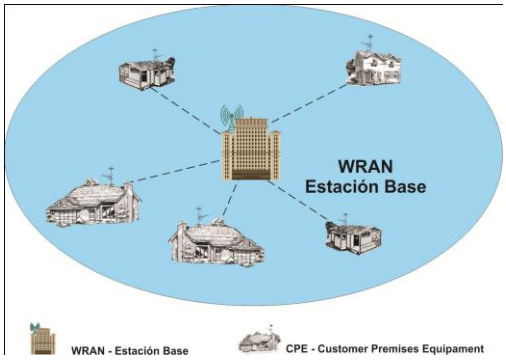


Figure 3. 802.22: Working scheme

Figure 3 illustrates a typical design. Assuming different quality of service (QoS), a Base Station - BS complying with the standard provides high-speed Internet services of up to 512 ***Customer Premise Equipments - CPEs***, fixed or portable, or groups of devices.

4.3. Cognitive Radio Capability

The cognitive radio capabilities supported by the standard are required to meet regulatory requirements for protection of frequency of incumbent's operators as well as to provide for efficient operation. They include: spectrum sensing, geolocation services, database access, registration and tracking of channel set management [8].

In areas where a computer with the IEEE 802.22 standard is intended to operate, the detection of operational channels which could be subject to interference includes the following:

- Television broadcasts.
- Wireless microphone transmissions.
- Transmissions from protecting devices, such as a Wireless Beacon[12].
- Other transmissions such as medical telemetry that may need to be protected in the local regulatory domain.

[11] ***IEEE 802.22.1: Standard to Enhance Harmful Interference Protection for Low-Power Licensed Devices Operating in the TV Broadcast Bands.*** 2010.
[12] IEEE 802.22.1.

4.4. Topology

The standard topology is point-to-multipoint. The protocol works in a master/slave procedure, so that each CPE requires approval from the BS to transmit.

The system functions with a ***Base Station - BS*** and multiple ***Customer Premise Equipment - CPEs***. The base station controls the whole link, as well as its own performance and the CPE stations. It executes media access control, modulation of the RF transmission, coding, and selection of operating frequencies.

The CPE uses an antenna system as shown in figure 4. It has a directional antenna similar to those used for transmitting/receiving TV signals, one sensing antenna that surveys the spectrum to determine which frequencies are available and a GPS antenna to determine the exact location of the transmitting station [8].

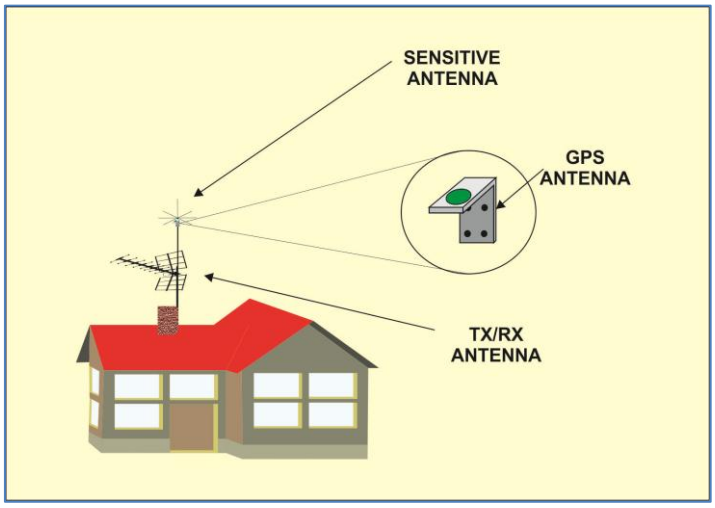


Figure 4. Customer Premise Equipment Antennas

When the sensing antenna detects a band of the spectrum in use, the cognitive radio system changes the transmission features to avoid interference while granting priority to TV operators.

The GPS determines the exact location of the detected signal, so that the system searches the database service of the regulatory agency and find free frequencies for frequency hopping. According to the received information, the base station changes or not the parameters of transmission/reception.

4.5. The IEEE 802 LAN/MAN Committee: Family of Wireless Standards

The ***IEEE 802 LAN/MAN Standard Committee*** has developed a large and diverse family of wireless data communication standards. Since the first 802.3 version to the present, they have dealt with different requirements in wireless communications.

Figure 5 illustrates the most significant wireless standards and the relative position of the 802.22.

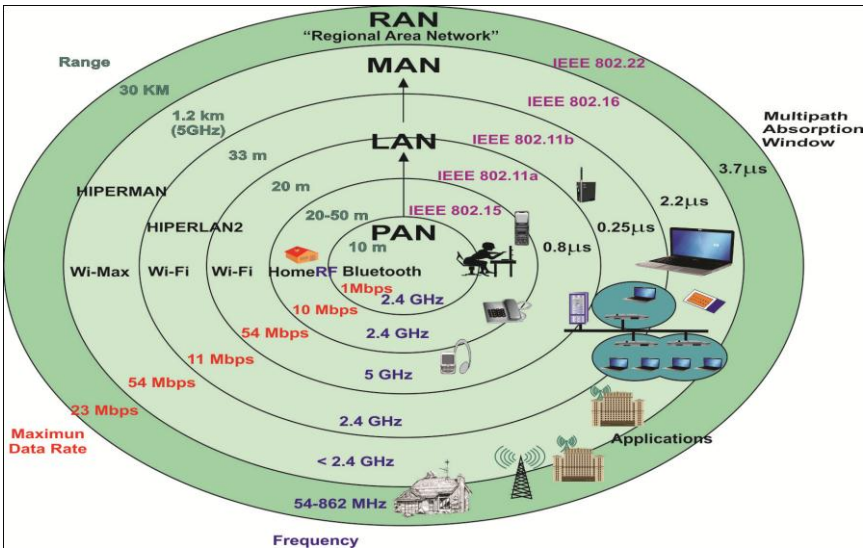


Figure 5. Different Wireless Standards Developed by the IEEE 802 Committee

The standard provides wireless broadband access in rural areas within a range of 30 up to a maximum of 100 km from a base station.

4.6. Physical Layer - PHY

Similarly to the ***Asymmetric Digital Subscriber Line – ADSL***, the IEEE 802.22 standard provides broadband access at a data transfer rate of 1.5 Mbps for uplink and 384 kbps for downlink (Figure 6).

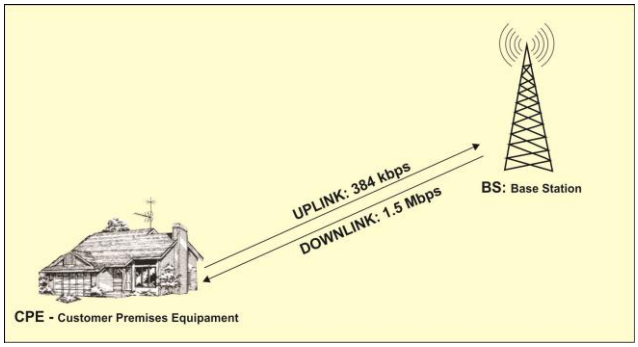


Figure 6. Different Wireless Standards Developed for the IEEE 802 Committee

It works with multiplexing ***Orthogonal Frequency Division Multiple Access - OFDMA*** and defines twelve combinations of three modulations: QPSK - Quaternary Phase Shift Keying, 16-QAM, and 64-QAM Quadrature Amplitude Modulation; and convolutional coding for error handling with the procedure ***Forward Error Control - FEC***.

Parameters	Specification
Frecuency range	54-862 MHz
Bandwidth	6 MHz, 7 Mhz, 8 Mhz
Payload modulation	QPSK, 12-QAM, 64-QAM
Transmit effective isotropic radiated power	Default 4 W for CPEs
Multiple access	OFDMA
Cyclic prefix modes	1/4, 1/8, 1/16, 1/32
Duplexing	TDD

Figure 7. Details the Different Features of the Standard

4.7. Medium Access Control Layer - MAC

The MAC layer supports cognitive capabilities. Thus, it must have mechanisms for flexible and efficient data transmission. It must guarantee the reliable protection of services in the TV band and should be allowed to coexist with other 802.22 systems.

This layer is applicable to any region in the world and does not require country-specific parameter sets.

It is ***connection-oriented*** and provides flexibility in terms of QoS support. It also regulates downstream medium access by TDM, while the upstream is managed by an OFDMA system. The BS manages all the activities within its cell and the associated CPEs are under the control of the BS.

5. Conclusions

Societies today have become highly dependent on the radio spectrum with the intensive use of wireless devices and communication services. Cognitive Radio, using intelligent software and taking advantage of white holes, may be a solution to spectrum saturation.

The Project Communitarian Private Networks has focused on evaluating solutions to the communication problems of rural areas. It has concluded that wireless communications may be among the feasible solutions.

Taking advantage that Argentina has a plan to cover a significant area of its territory with a TV broadcasting system, the conditions may be the ideal to introduce simultaneously the 802.22 standard to the problem of rural communications.

The Project ***Communitarian Private Networks*** continues its work on this line of research.

6. Acknowledgements

The financial support provided by Agencia Nacional para la Promoción Científica y Tecnológica (Project PICTO 11- PICTO 11-18621 is gratefully acknowledged.

7. References

[1] Antonio Castro Lechtaler (Director). PICTO 11-18621. Redes Privadas Comunitarias. Proyecto FONCyT, ANPCyT. Working Paper.

[2] J. Garcia Guibout, C. García Garino, A. Castro Lechtaler, R. Fusario and Guillermo Sevilla. Physical and Link Layer in Power Line Communications Technologies. Proceedings of 13th of Argentine Congress on Computer Science. ISBN 978 - 950 – 656 – 109 – 3. pp. 56 a 67. Corrientes. October 2007.

[3] J. García Guibout, C. García Garino, A. Castro Lechtaler, R. Fusario and Guillermo Sevilla. Power Line Communications in the Electric Network. Proceedings of 13th of Argentine Congress on Computer Science ISBN 978 - 950 – 656 – 109 – 3. pp. 68 a 79. Corrientes. October 2007.

[4] J. García Guibout, C. García Garino, A. Castro Lechtaler and R. Fusario. Transmission voice over 802.11. Proceedings of 14th of Argentine Congress on Computer Science. ISBN 978 - 987 - 24611 - 0 - 2. pp. 307 a 318. Chilecito. October 2008.

[5] A. Castro Lechtaler, A. Foti, R. Fusario, C. García Garino and J. García Guibout. Communication Access to Small and Remote Communities: The Corral de Lorca Project. Proceedings of 15th of Argentine Congress on Computer Science. ISBN 978 - 897 - 24068 - 4 - 1. pp. 1.117 a 1.126. Jujuy. October 2009.

[6] A. Castro Lechtaler, A. Foti, C. García Garino, J. García Guibout, R. Fusario and A. Arroyo Arzubi. Proyecto Corral de Lorca: Una solución de conectividad a grupos poblacionales pequeños, aislados y distantes de centros urbanos. Proceedings de la Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI 2010. - Volume III - ISBN - 13: 978 – 1 – 934272 – 96 - 1. PP. 121 a 127. Orlando, USA. June 2010.

[7] http://www.cplus.org/rmw/index.html (Radio mobile software).

[8] IEEE 802.22 - Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Policies and Procedures for Operation in the TV Bands.

[9] Carlos Cordeiro, Kiran Challapali, and Dagnachew Birru, Sai Shankar N. IEEE 802.22: An Introduction to the First Wireless Standard based on Cognitive Radios Journal of Communications, Vol. 1, N° 1, april 2006.

[10] C. Gómez. Spectrum Regulation and Policy Officer Radiocommunication Bureau. www.itu.int/ITU-D/asp/CMS/Events/.../ITU-APT-S3_**Cristian_Gomez**.pdf. ITU. Apia, Samoa. April 2013.

[11] CEPT Report 24. A preliminary assessment of the feasibility of fitting new/future applications/services into non-harmonized spectrum of the digital dividend (namely the so-called "***white spaces***" between allotments. Report C from CEPT to the European Commission in response to the Mandate on: Technical considerations regarding harmonization options for the Digital Dividend. 1 July 2008.

[12] http://www.wirelessinnovation.org/introduction_to_sdr

[13] Dillinger, M; Madani, K; Alonistioti, N. Software Defined Radio: Architectures, Systems and Functions. Ed. Wiley & Sons, 2003.

[14] J. Mitola, G. Maguire. Cognitive radio: making software radios more personal. IEEE Personal Communications Magazine, vol. 6, Nr. 4, pp. 13–18, Aug. 1999.

[15] J. Mitola. Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio. Dissertation submitted in partial fulfillment of the degree of Doctor of Technology. Royal Institute of Technology (KTH) - Teleinformatics. ISSN 1403 - 5286. Sweden. May 8, 2000.

Posicionamiento indoor determinado por la distancia en función de la potencia medida de balizas bluetooth

Marcelo MARINELLI[1], Juan TOLOZA[2, 3], Nelson ACOSTA[2]

[1]Departamento de Informática, Facultad de Ciencias Exactas Químicas y Naturales.
Universidad Nacional de Misiones
[2]Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Provincia de Buenos Aires
[3]Becario postdoctoral CONICET
marcelomarinelli@gmail.com, {jmtoloza, nacosta}@exa.unicen.edu.ar

Abstract. El presente artículo presenta una experiencia de captura de datos de dispositivos que emiten señales bluetooth usando una placa tipo Arduino Mega 2560. Se analizan los datos con algunas técnicas y se detalla la magnitud de errores encontrados en las diferentes muestras. Además, se proponen algunas nuevas medidas para intentar alcanzar una mejor precisión en el posicionamiento.

Keywords: Posicionamiento indoor, Balizas Bluetooth, RSSI.

1 Introducción

Los métodos de posicionamiento fueron evolucionando en el tiempo. Los fenicios usaban el sol, la luna y las estrellas para guiarse [1][2]. En la actualidad, se usan micro dispositivos electrónicos [3]. A finales del siglo XX, la aparición de calculadoras y computadoras electrónicas, facilitó grandemente el cálculo; pero la aparición del GPS, poco después, revolucionó la forma de localizar un objeto [4].

Esta tecnología para el posicionamiento en exteriores carece de utilidad en un espacio cerrado como puede ser un edifico. Es difícil usar esta tecnología para distinguir en que habitación o en que planta se encuentra ubicado una persona. Es por ello, que en los últimos años se han presentado diversas soluciones de posicionamiento en espacios interiores. Entre ellas, se encuentra Bluetooth la cual se experimenta aquí y se analiza con diversas técnicas para lograr posicionar un objeto en un ambiente interior.

Bluetooth utiliza frecuencias de radio del orden de 2.4 Ghz, representa una tecnología económica [5], pero es de corto alcance, de esta manera, para cubrir la zona de un recinto, se necesitarían varios dispositivos. El error asociado a la estimación puede encontrarse en torno a los 1.5 metros de precisión. En este sentido, el parámetro RSSI Received Signal Strength Indicator, (Intensidad de la Señal Recibida) no es preciso, motivo por el cual, no se puede estimar con exactitud la

ubicación de un dispositivo, sino que se identifica el entorno en el que se encuentra, en un radio determinado.

En general, los sistemas de posicionamiento heredan características dependientes del tipo de sensor. Algunas de ellas son: Retardo en la propagación, difracción, reflexión y la dispersión que afectan a todos los tipos de señales. Las características propias de la señal son las siguientes [6]:

- Atenuación por distancia: A mayor separación entre el emisor y el receptor, la potencia de la señal decrece con el tiempo de forma logarítmica, si el receptor se encuentra a una distancia corta del emisor, la potencia decrece rápidamente y si el receptor se encuentra en un rango de alcance medio, la señal decrece a una velocidad menor.
- Absorción de la señal: Cuando la señal atraviesa algún material, la potencia de la misma se debilita o atenúa en mayor o menor intensidad, dependiendo de las características físicas del material y de la frecuencia propia de la onda.
- Reflexión: Este fenómeno ocurre, cuando una onda, choca con un obstáculo, parte de la potencia de la señal no se absorbe, sino que es reflejada y la misma puede tener distinta fase que la señal original, dependiendo de las características propias del obstáculo.
- Dispersión: Este fenómeno ocasiona que parte de la energía sea irradiada en numerosas direcciones diferentes y ocurre cuando el medio por el cual viaja la señal, está formado por objetos con dimensiones pequeñas, comparados con la longitud de onda propia de la señal. Es el fenómeno contrario a la reflexión, la cual ocurre cuando los objetos poseen dimensiones grandes.
- Difracción: Cuando una señal impacta con el borde de un obstáculo, se originan diferentes frentes de onda en distintas direcciones. Los factores de los cuales dependen la intensidad de este fenómeno son: la calidad y tipo del material con el que está compuesto el obstáculo, así como también de la amplitud o fase de onda.

Los tres últimos fenómenos citados anteriormente, dan lugar un fenómeno denominado: Multitrayecto (Multipath) que origina que la señal llegue al receptor a través de diferentes caminos y por lo tanto a diferentes tiempos ocasionando retardos e interferencias en las transmisiones. De esta manera, las comunicaciones inalámbricas en interiores se caracterizan por este fenómeno, donde no solamente existen señales directas entre el emisor y el receptor, sino que también se encuentran señales difractadas, dispersadas y reflejadas por los diferentes obstáculos y objetos que se encuentran en el medio.

En el trabajo desarrollado en [7] se dispone de tres ordenadores portátiles que actúan como emisores de señal Bluetooth y un dispositivo móvil como receptor de la señal Bluetooth. El dispositivo móvil es quien calcula la posición donde se encuentra mediante triangulación de la señal que emiten los tres portátiles.

En [8] se presenta una arquitectura en la que los emisores son de bajo costo y el receptor es un dispositivo móvil. El artículo quiere enfatizar en la ventaja de usar este tipo de arquitectura pasiva de bajo costo.

La plataforma Alipe [9] mezcla diversas topologías para obtener la posición. En esta plataforma por un lado hay dispositivos Bluetooth que envían información sobre su ubicación al realizarle una petición por parte de otro dispositivo Bluetooth cliente. Si el dispositivo al que se le ha realizado la petición no está adaptado para comunicar su posición, el dispositivo cliente que ha realizado la petición registra la dirección Bluetooth remota y busca en una base de datos centralizada la ubicación asociada a esa dirección Bluetooth.

"Follow me" [10] presenta una aplicación práctica para un sistema de posicionamiento en interiores. El sistema consiste en dispositivos Bluetooth rastreadores cuya ubicación es conocida. Estos dispositivos rastreadores están constantemente escaneando dispositivos Bluetooth, cada vez que se detecta un dispositivo se almacena su dirección Bluetooth junto con su ubicación, que es la ubicación del dispositivo rastreador, en una base de datos centralizada. La aplicación ofrecida al usuario es poder obtener su ubicación en el edificio, consultando esa base de datos y publicar la ubicación en una aplicación web como puede ser Twitter.

En el pabellón de Finlandia en la expo de Shangai 2010 se desarrolló una aplicación para móvil [11] en la que los usuarios podían obtener su posición dentro del pabellón mediante puntos de acceso Bluetooth que calculaban posiciones mediante distribuciones de probabilidad del indicador RSSI.

Los sistemas de posicionamiento en interiores no sólo se limitan el uso de tecnologías inalámbricas también, como se puede ver en [12], se ha desarrollado un sistema de posicionamiento en interiores basado en visión por ordenador, en el que se cuenta con una base de datos de imágenes georefenciadas que se usa para buscar coincidencias con lo que está viendo el dispositivo móvil en ese momento.

En la mayoría de los sistemas de posicionamiento indoor analizados se opta por estimar la posición usando el indicador de Fuerza de Señal de Recepción (RSSI), recolectando medidas desde distintos puntos para inferir un modelo probabilístico que estima las posiciones una vez que el sistema está en funcionamiento.

Albert Huang [13] utiliza la infraestructura de computadoras existentes en un edificio agregando 30 balizas BT en los puertos USB de las mismas, logrando una exactitud de 3-10 metros, dependiendo de la densidad de baliza en la zona.

Fernández Gorroño [14] utiliza un sistema de posicionamiento en interiores basado en la tecnología Bluetooth para aplicaciones en Dispositivos Móviles (DM), con el objeto de proveer información en estaciones de subterráneos. La característica de este sistema es que el peso de la lógica está en el DM y las balizas son pasivas y de bajo costo.

Sudarshan S. Chawathe [15] estudia los retrasos en la fase de descubrimiento de balizas Bluetooth y propone métodos para aliviarlos de forma de poder utilizarlos para aplicaciones de localización en interiores como, complejos de edificios grandes o terminales de aeropuertos usando coordenadas adecuadas al lugar como número de piso y número de habitación o terminal del aeropuerto y dársena.

La sección 2 presenta el proceso de adquisición de datos, la 3 muestra el análisis que se hace a los datos recavados con algunas técnicas propuestas, aquí se muestran los resultados a los que se llegó luego de procesar los datos. Por último, la 4, presenta las conclusiones y futuros trabajos.

2 Toma de muestras y experimentos realizados

Se diseña un escenario para la toma de muestras en una de las oficinas del edificio INTIA/INCA, de la Facultad de Ciencias Exactas perteneciente a la Universidad Nacional del Centro de la Provincia de Buenos Aires. Se hacen marcas de precisión en el suelo cada un metro llegando a completar 25 metros. Se toman 100 muestras de cada punto. Las primeras marcas, hasta los 3 metros corresponden al interior de una oficina y hasta los 7 metros a otra oficina contigua. Luego hasta los 25 metros las tomas se hacen en un pasillo del edificio.

Para la toma de datos se especifica un sistema de medición de potencia de dispositivos Bluetooth (BT) remotos (baliza BT) utilizando comandos AT. Para ello, se desarrolla un programa de control residente en la placa Arduino Mega que setea al dispositivo Bluetooth local en modo "master". De esta forma, se interroga a los dispositivos detectados, se identifican por su dirección MAC y se mide su potencia. La figura 1 muestra un diagrama de los componentes del sistema de adquisición.

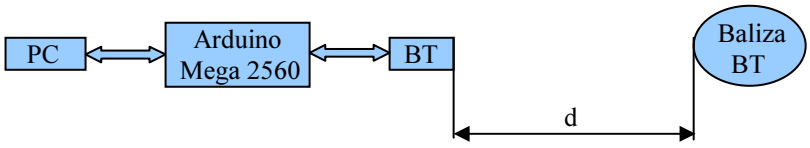


Figura 1. Diagrama de componentes del sistema

Los dispositivos que se usan para adquirir los datos son módulos transceptores de tecnología inalámbrica Bluetooth RS232 TTL V2.0 con chipset RSE.

Para el procesamiento de los datos se desarrolla un programa en Ansi C que, por medio del puerto USB conectado al Arduino, recibe cien datos de medición de potencia y los almacena en un archivo de texto plano.

3 Análisis de los datos con las técnicas propuestas

Los datos obtenidos se vuelcan en una plantilla de cálculo donde se obtienen por cada grupo de datos: la moda, el valor máximo, el mínimo y el promedio. Luego se proponen algunas técnicas novedosas a implementar aplicadas en [16][17][18].

En la figura 2 se observa el comportamiento de cada una de las técnicas aplicadas. Cabe aclarar que el mínimo no se muestra en el gráfico ya que tiene valores extremos que no permiten visualizar con claridad el resto de los resultados obtenidos.

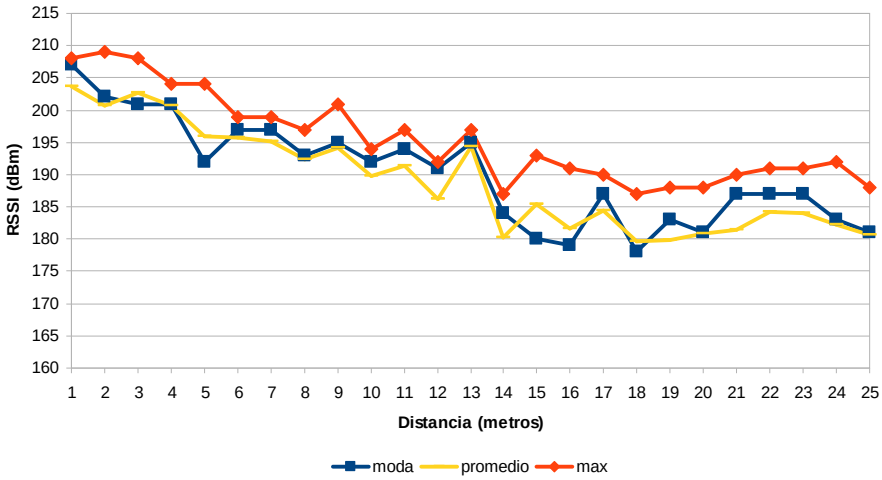


Figura 2. Gráfico de la potencia en función de la distancia

La figura 2 presenta un comportamiento similar para las tres medidas. Cuando la distancia aumenta, el indicador de fuerza de la señal (RSSI) disminuye. Hasta los 3 metros puede verse como el comportamiento es ideal para la moda ya que cada uno de los valores es menor a medida que se aleja de la baliza. No pasa lo mismo con las otras medidas. Al pasar a la oficina contigua, y esto puede ser debido a la presencia de paredes que obstaculizan la señal, la moda oscila bruscamente pero el promedio y los máximos mantienen consistencia hasta los 8 metros inclusive. Desde allí y hasta los 12 metros, el comportamiento es similar para todas las medidas. A los 13 y 14 metros, prácticamente coinciden en valor las tres. Luego, se observan muchas oscilaciones en todas las medidas, esto se puede atribuir a los rebotes de la señal en las paredes del pasillo y en el amoblamiento de las oficinas que se encuentra al paso de la señal.

Ahora si se calcula la regresión lineal de cada una de las medidas se puede observar la distancia entre la medida “ideal” y la real. En el caso de la figura 3, se observa la regresión lineal para el promedio, en la 4 para la moda; y finalmente en la 5, para el máximo. También se muestran las ecuaciones de las rectas en cada caso y el valor de R^2 que representa cuanto se ajusta la recta a los valores reales obtenidos. Cuanto mayor es este valor, significa que mejor se ajusta la regresión a los valores reales.

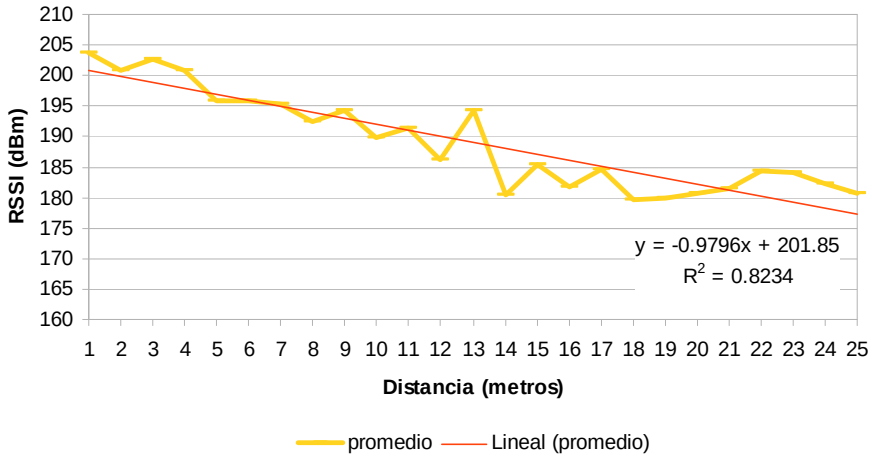


Figura 3. Promedio y su regresión lineal

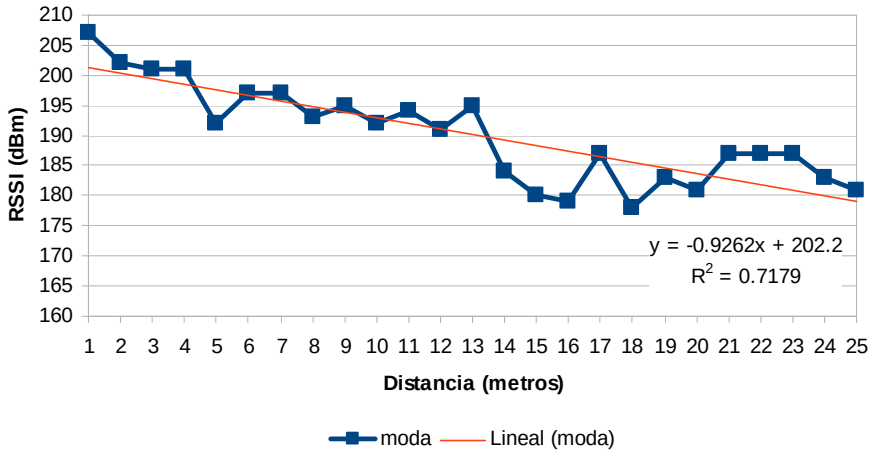


Figura 4. Moda y su regresión lineal

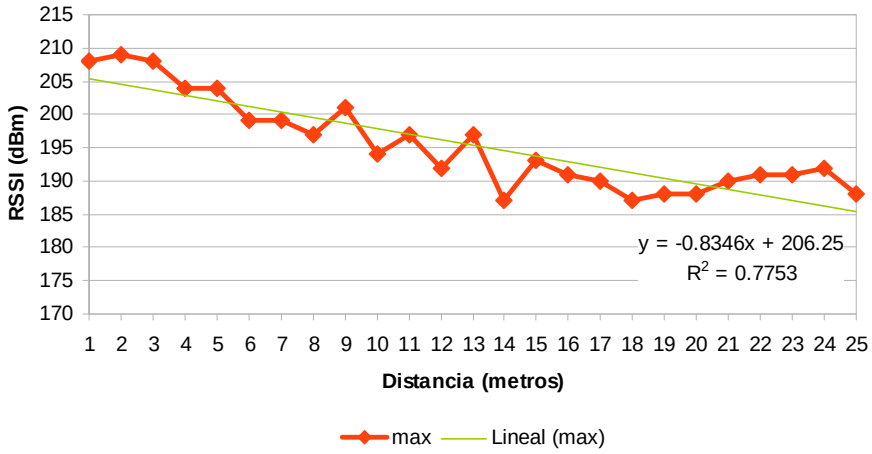


Figura 5. Máximo y su regresión lineal

Según lo analizado la mejor regresión, es decir, el mayor R^2 se da para el promedio. Por lo tanto, el mejor estimador estadístico para el caso estudiado que más se asemeja a un comportamiento lineal es el promedio. Entonces si se aplica a los datos de entrada (RSSI) la ecuación x = (y - 201.85) / -0.9796 que se despeja de la que se observa para la regresión (y = -0.9796x + 201.85), se puede tener una salida (metros) donde se obtenga una distancia a la baliza de tal manera de estimar su posición relativa.

Las siguientes tres figuras (6, 7 y 8) analizan la magnitud del error cometido entre la medida tomada y la que indicaría la regresión lineal. Esto se realiza luego de aplicar la ecuación despejada para la entrada (x) como se mostró en el caso del promedio. En la figura 6 se muestra el error del promedio, en la 7 de la moda y en la 8 del máximo.

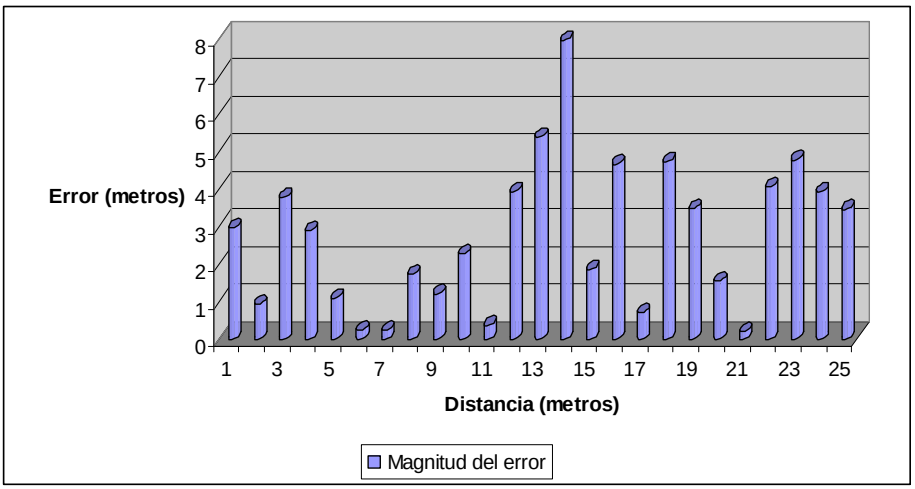


Figura 6. Magnitud del error entre el promedio y su regresión lineal

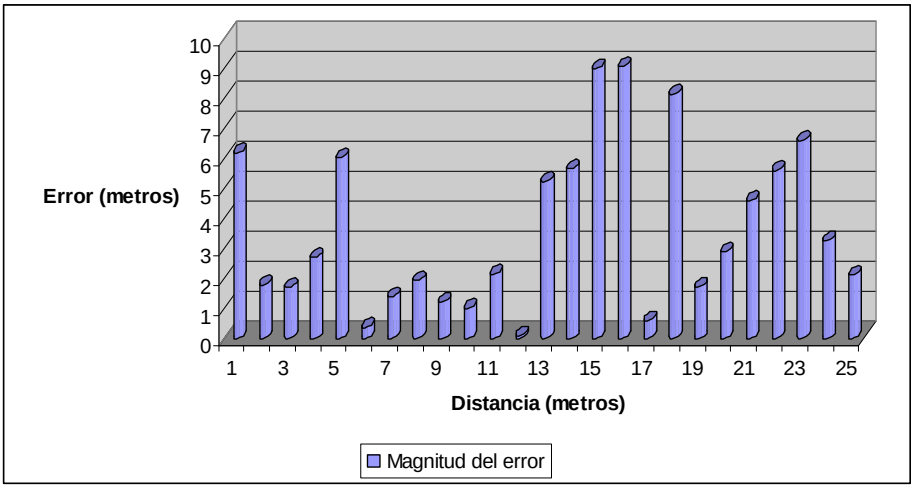


Figura 7. Magnitud del error entre la moda y su regresión lineal

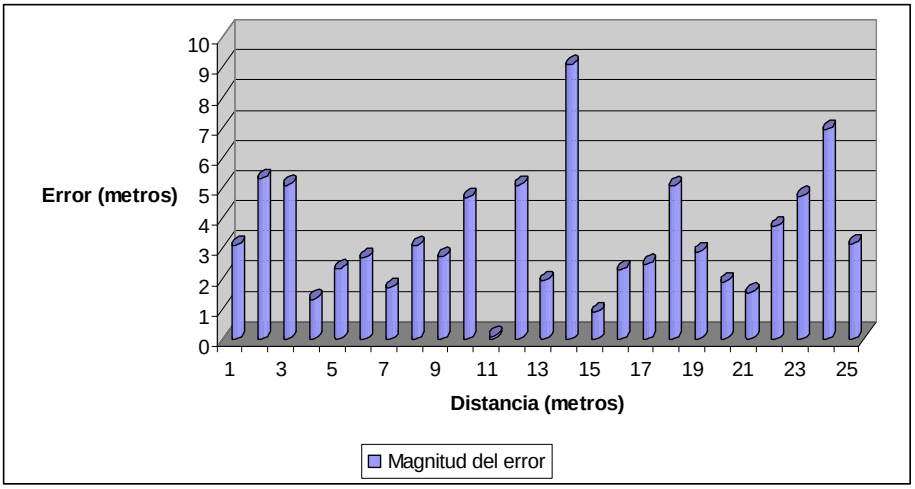


Figura 8. Magnitud del error entre el máximo y su regresión lineal

El error máximo para todos los casos está cercano a los 9 metros cuando la distancia al dispositivo es de 14 metros aproximadamente. El mínimo es 0 para algunos casos. En las tres estimaciones se denota que en el rango medio de distancia es donde se obtiene la mayor magnitud de error. Este fenómeno se puede deber a los obstáculos que debe sortear la señal para llegar al dispositivo y obviamente cuando más se aleja, menor es la recepción. También se observa que cercano al dispositivo de captura el error medido es alto pero no se tienen detalles de su ocurrencia.

4 Conclusiones y trabajos futuros

Se ha presentado un sistema ad-hoc para posicionamiento indoor usando un entorno libre como es Arduino con la posibilidad de ser montado en cualquier móvil. Las técnicas utilizadas muestran que es posible estimar la distancia con un cierto grado de error y por medio de triangulación, localizar un objeto en un ambiente interior. Los errores máximos se encuentran en los 9 metros cuando la distancia es casi del doble. Calculando el promedio de las medidas se obtiene un buen estimador pero no es suficiente para una buena precisión. La moda si bien es la que menos se ajusta a su regresión lineal, también es un buen estimador aunque el error acumulado es el más grande de los tres.

Como futuros trabajos se prevé el desarrollo de técnicas que permitan mejorar la estimación de la distancia a partir de la fuerza de la señal recibida. Se va a tomar mas cantidad de muestras con rangos de distancias que van de a 50cm para verificar la aplicabilidad de las técnicas propuestas en este trabajo y de las futuras a desarrollar. También se va a utilizar el tiempo de vuelo de la señal para verificar los datos obtenidos. Se van a hacer pruebas en un ambiente sin obstáculos para ver el comportamiento de la señal en un ambiente ideal. Además, se van a realizar tomas de datos en un ambiente exterior de manera de analizar su comportamiento.

Como se mencionó anteriormente, algunas de las técnicas que se proponen implementar son: filtro de Kalman con ajuste de la desviación estándar, lógica difusa y redes neuronales. Con ello se busca aumentar la precisión de posicionamiento.

Referencias

1. Rao (2010) Global Navigation Satellite Systems. Tata McGraw-Hill Education, 478 pp.
2. Misra P. & Enge P. (2010) Global Positioning System: Signals, Measurements, and Performance. New York, Ganhga-Jamuna Press, 590 pp.
3. Asdrúbal V. (2004) De la técnica a la modernidad: Construcciones técnicas, ciencia, tecnología y modernidad. Universidad de Antioquia, 263 pp.
4. Maini A. K. & Agrawal V. (2010) Satellite Technology: Principles and Applications. 2nd Edition, John Wiley & Sons, United Kingdom, 704 pp.
5. Hallberg J., Nilsson M., Synnes K. (2003) Bluetooth Posittioning. 10th International Conference on Telecommunications, Volume 2, IEEE, pp. 954-958.
6. Stewart J. W. (1983) Introduction to Wave Propagation, Transmission Lines, and Antennas. Navy Electricity and Electronics Training Series. U.S. Navy.
7. S. Feldmann, K. Kyamakya, A. Zapater, Z. Lue, An Indoor Bluetooth-Based Positioning System: Concept, Implementation and Experimental Evaluation, in: International Conference on Wireless Networks, 2003.
8. Cheung K., Intille S., and Larson K., 2006. An Inexpensive Bluetooth-Based Indoor Positioning Hack. Proceedings of UbiComp 2006
9. J. Hallberg, M. Nilsson, K. Synnes, “Bluetooth Positioning”, The Third Annual Symposium on Computer Science and Electrical Engineering (CSEE 2002), Luleå, Sweden, 27-28 May 2002
10. Polychronis Ypodimatopoulos and Andrew Lippman. 'Follow me': a web-based, location-sharing architecture for large, indoor environments. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 1375-1378.

11. Pei, Ling; Chen, Ruizhi; Liu, Jingbin; Tenhunen, Tomi; Kuusniemi, Heidi; Chen, Yuwei; , "An Inquiry-based Bluetooth indoor positioning approach for the Finnish pavilion at Shanghai World Expo 2010," Position Location and Navigation Symposium (PLANS), 2010 IEEE/ION , vol., no., pp.1002-1009, 4-6 May 2010
12. Paucher, R.; Turk, M.; , "Location-based augmented reality on mobile phones," Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on , vol., no., pp.9-16, 13-18 June 20
13. Huang, A., "An Inexpensive Bluetooth-Based Indoor Positioning Hack MIT" CSAIL. 2005. Available at http://people.csail.mit.edu/albert/pubs/2005-ashuang-sm-thesis.pdf
14. Fernández Gorroño, J. L. SISTEMA DE GUIADO MULTIMEDIA EN INTERIORES MEDIANTE DISPOSITIVOS MÓVILES BLUETOOTH.
15. Chawathe, S.S., "Low-latency indoor localization using bluetooth beacons," Intelligent Transportation Systems, 2009. ITSC '09. 12th International IEEE Conference on , vol., no., pp.1,7, 4-7 Oct. 2009
16. Toloza J. (2013) Algoritmos y técnicas de tiempo real para el incremento de la precisión posicional relativa usando receptores GPS estándar. Tesis Doctoral, Universidad Nacional de La Plata, 213 pp.
17. Toloza J., Acosta N. & De Giusti A. (2012) "An approach to determine the magnitude and direction error in GPS system," Asian Journal of Computer Science and Information Technology, Volume 2 No. 9, pp. 267-271.
18. Acosta N. & Toloza J. (2012) "Techniques to improve the GPS precision," International Journal of Advanced Computer Science and Applications, Volume 3 No. 8, pp. 125-130.

Posicionamiento WIFI con variaciones de Fingerprint

Carlos KORNUTA[1], Nelson ACOSTA, Juan TOLOZA[2]

INCA/INTIA - Facultad de Ciencias Exactas - UNICEN
TANDIL - Argentina
{ ckornuta, nacosta, jmtoloza }@exa.unicen.edu.ar

Abstract. Los sistemas de posicionamiento Indoor estiman la posición de un dispositivo móvil en un entorno cerrado con una precisión relativa. Existen diversas técnicas de posicionamiento, donde el parámetro mayormente utilizado es el RSSI (*Received Signal Strength Indicator*). En este artículo se analiza la técnica *Fingerprint* con la finalidad de estimar el margen de error obtenido con la distancia euclidiana como métrica principal. Se presentan variantes de la construcción de la base de datos *Fingerprint* analizando distintos valores estadísticos con la finalidad de comparar la precisión de diferentes indicadores.

Keywords: Posicionamiento indoor, Localización indoor, RSSI, Fingerprint

1. Introducción

En la actualidad, es necesario contar con mecanismos que posibiliten determinar la ubicación de un dispositivo móvil en el interior de un edificio. Algunos ejemplos de ellos son mapas interactivos de centros comerciales y museos, mapas guiados de campus universitarios, sistemas de monitorización de pacientes en hospitales y/o albergues de personas mayores [1]. Se descarta el uso de GPS, ya que, no puede ser utilizado en ambientes interiores, porque necesitan una línea de visión clara y sin obstáculos entre el dispositivo y un mínimo de tres satélites [2] [14].

La estimación de la posición relativa de un dispositivo móvil, en adelante DM, es el proceso mediante el cual se obtiene información sobre la posición, con respecto a referencias sobre un espacio predefinido [3]. Las técnicas para estimación de la posición Indoor, dependiendo la tecnología de sensores utilizada, son: Tiempo de Arribo (ToA), Ángulo de Arribo (AoA), Indicador de potencia de la señal (RSSI). Dentro de esta última, el algoritmo más utilizado para estimar la posición es *Fingerprint* [4].

En el año 2000, el sistema RADAR [5] obtiene una precisión media en el rango de 2-3 m. En 2003 el sistema LEASE [6] consigue una precisión de 2.1 m. En 2007, se utiliza la técnica de *Fingerprint* [7] para estimar la posición del DM en conjunto con un algoritmo de redes Bayesianas, logrando una precisión de 1.5 m. En [8] se presenta un sistema que

[1] Becario CONICET Tipo I
[2] Becario Postdoctoral CONICET

utiliza *Fingerprint* y el método de la distancia euclidiana con un algoritmo de mejora utilizando lógica difusa, en una primera instancia obtienen una precisión de 4.47 m y luego con lógica difusa 3 m. El sistema de posicionamiento EKAHAU [9] basado en el parámetro RSSI logra una precisión de 1-5 m dependiendo de las condiciones del entorno. En [10] se presenta un sistema basado en un algoritmo utilizando redes neuronales, logrando la precisión de 1-3 m. En este artículo se analizan diversas variantes de la construcción de la base de datos *Fingerprint*.

La sección 2 muestra el funcionamiento de la localización usando *Fingerprint* y plantea nuestra propuesta, la 3 muestra la experimentación, en la 4 se analizan los datos, la 5 muestra el análisis de los datos y la 6 las conclusiones y futuros trabajos.

2. Localización utilizando FINGERPRINT

El método de *Fingerprint* se basa en que cada posición dentro de un recinto tiene una única firma, compuesta por una tupla (P/L), en donde P contiene información acerca del patrón único y L información relativa a la posición dentro del edificio. La información relativa a la posición, puede ser representada en un formato de tupla de coordenadas o una variable representativa. Esta técnica requiere entrenamiento, donde se realiza el muestreo de cada una de las firmas [11].

En primer lugar, se debe diseñar un *Radio Map* [6], que es un mapa patrón conteniendo las posiciones especificas dentro del edificio y un vector de potencias RSSI que contiene todas las potencias de los Access Point, en adelante AP, alcanzados en cada posición. La creación de un *Radio Map* incluye:

1. En cada posición del escenario se muestrean los valores de potencia de señal (RSSI), armando un vector de potencias para cada posición, cuya dimensión depende de la cantidad de AP visibles.
2. Para cada sector del área que puede recibir señal de N puntos de acceso, se obtiene un vector de los RSSI de cada AP.
3. Para vincular la firma y la información de localización se utiliza un método determinístico, para encontrar la posición del vector más cercano, en muchos casos se usa la distancia Euclídeana.

Para estimar la posición del DM, se capturan los valores de todos los AP visibles desde la posición que se quiere estimar. Los valores adquiridos son comparados con los valores obtenidos en el *Radio Map* para obtener las coordenadas de ubicación del dispositivo [12].

La base de datos *Fingerprint* es un resumen de los datos de la *Radio Map*, que facilita la ubicación minimizando el cálculo y reduce el error. Los algoritmos de estimación correlacionan los valores obtenidos entre la información de la localización y la base *Fingerprint*, para determinar la posición relativa del DM. El método determinístico más conocido es el "*vecino más próximo*", donde se utilizan los vectores medios, los cuales contienen el promedio de los valores RSSI de cada AP en cada punto del mapa.

3. Creación del Radio Map y la base de datos Fingerprint

La experimentación se realiza en el sector de becarios del instituto de investigación INTIA/INCA, de la Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires. El área tiene una dimensión aproximada de 36 m^2. Para realizar las mediciones y captura de datos se divide el área correspondiente en un eje de coordenadas (*fila*, *columna*) (Figura 1), cada región del mapa tiene una separación de 90 cm con respecto al punto anterior.

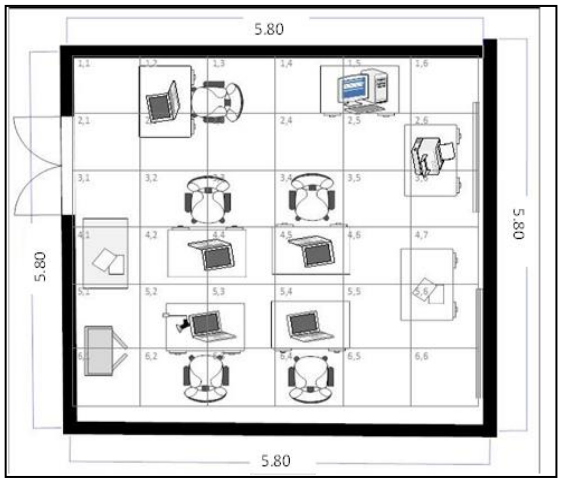


Figura 1. Ubicación de las coordenadas en el mapa

La captura de datos se realiza usando IWLIST en Ubuntu 8.04. El proceso de captura de datos para el armado del *Radio Map* (Figura 1) es:

1. Posicionamiento del DM en un punto de coordenadas del mapa.
2. Escaneo y captura de RSSI por 180 segundos para estabilizar la señal.
3. Escaneo, captura y almacenamiento de RSSI y SSID, de la señal de los diferentes AP que están al alcance, en esa posición por 90 segundos.
4. Traslado del dispositivo al siguiente punto de coordenadas del mapa, y se retoma en el punto (1) si no es el último.

En la Figura 2 se visualiza la distribución de los AP. El sector de becarios del INTIA es el AP 4, y el edificio tiene 58 m de largo. Por cada punto de muestreo se obtienen 100 vectores de parámetros RSSI, conteniendo 11 valores correspondientes a los AP disponibles en el rango de alcance del DM. Por convención, cuando un AP se encuentra fuera del rango de alcance, se asigna el valor 0.

Con los datos del *Radio Map* se construye la base de datos *Fingerprint* (formada por los valores promedio), y además se ha estudiado otros valores para representar cada AP en cada posición:

- **Media RSSI:** la media aritmética de todas las observaciones del AP.
- **Dupla Intercuartílica:** Considerando el total de valores obtenidos por cada AP, se ordenan los datos en forma ascendente, luego se divide en 4 conjuntos con igual cantidad de elementos. Se eliminan los cuartiles extremos, y de los cuartiles centrales se calcula:
    - promedio y
    - moda aritmética.
- **Moda:** Se calcula la moda con el total de muestras obtenidas.
- **Promedio de Dupla Intercuartílica:** Se promedian los valores promedio y moda de la Dupla Intercuartílica.

Figura 2. Distribución de los AP, denominados 1: chacra, 2: default, 3: inca, 4: inca2, 5: intia, 6: isistan-2, 7: plademа-2, 8: plademа-invitado, 9: slab, 10:unicen2, 11: wlbiolab.

4. Análisis de datos

Tomando como referencia el AP 3, que se encuentra a aproximadamente 15 m, 4 paredes de ladrillos y un durlock, del lugar donde se realiza la captura y muestreo de los valores correspondientes a la señal de los AP encontrados. Las variaciones con respecto a la potencia de la señal, analizando por filas, son las siguientes:

- Desde la posición inicial (1,1) en línea recta, cada 180 cm la potencia de la señal, disminuye en 2 dBm. En la posición (1,5), vuelve a su valor normal y luego vuelve a disminuir. Por lo que fluctúa entre -86 y -90.
- Si consideramos la fila 2 de coordenadas, el valor de la potencia de la señal, disminuye luego de los 270 cm en 2 dBm.
- La fila 3, el valor disminuye, en 360 cm en 2 dBm, y aumenta a -89 en el punto siguiente, para volver a su punto inicial en el siguiente par de coordenadas.
- La fila 4, el valor se mantiene constante, sin grandes variaciones, hasta el último punto, en el cual el valor de la potencia aumenta a -79.

- La fila 5, desde la posición inicial, luego de 90 cm, el valor de la señal disminuye a -86, se mantiene estable y en el punto (5,4), vuelve a aumentar la potencia en 3 dBm y disminuye la misma a -92, y se mantiene estable en los valores iniciales.
- La fila 6, luego de 180 cm la señal aumenta 5 dBm y disminuye, alcanzando el valor -91, y regresa a los valores iniciales.

En contraste con los valores obtenidos por el AP 4 que se encuentra dentro del mismo sector de muestreo. Los valores de la potencia de la señal, son los siguientes:

- Se comienza (1,1) con un valor inicial de -54, luego la señal fluctúa entre +- 9 dBm, a excepción del punto (1,6).
- La fila 2, existen fluctuaciones y variaciones menores que en el punto anterior, la señal oscila en un rango de 3 dBm, con la excepción del punto (2,5) que la señal disminuye 10 dBm y culmina con un valor aproximado al -55 dBm.
- La fila 3, varía entre -41 y -17 dBm, se estabiliza cerca de sus valores iniciales.
- La fila 4, varía entre -43 y -8 dBm en (4,3), en la posición (4,5) a 180 cm vuelve a estabilizarse en sus valores iniciales.
- La fila 5, varía entre -47 y -55 dBm, y entre -47 y -55 dBm, oscilando en 8 dBm
- La fila 6, varía entre -49 y -59 dBm, con una variación de +-10 dBm.

Se infiere al efectuar un análisis de los datos que los valores de la señal fluctúan en un espectro más amplio, si el AP se encuentra a menor distancia física del punto de muestreo. Si el AP se encuentra más distante del punto de referencia, el valor de la señal no tiene grandes cambios, oscila en +-3 aproximadamente.

En la Tabla 1 se presentan los valores de absorción de la señal Wifi según el Material [13], que influyen en la degradación del parámetro RSSI.

Tabla 1. Atenuación de la potencia Wifi producida por los materiales a 2.4 GHz:

Obstáculo	Pérdida Adicional (dB) (aprox.)
Ventana no metálica (Vidrios)	3
Ventana Metálica	5 a 8
Pared Fina	5 a 8
Pared Media	10
Pared Gruesa (15 cm)	15 a 20
Pared muy gruesa (30 cm)	20 a 25
Piso o techo grueso	15 a 20
Piso o techo muy grueso	15 a 25

En la Tabla 2 se identifican 4 coordenadas principales dentro del mapa que corresponden a puntos determinados en los cuales podría existir una discrepancia de los valores y del conjunto de AP detectados, se seleccionan tres AP a modo ejemplo, identificando RSSI promedio, máximo y mínimo.

Tabla 2. Análisis de la variación de los AP

Coordenadas	AP	Promedio	Máximo	Mínimo
1.1	3	-89	-81	-97
	7	-75	-71	-79
	6	-64	-57	-69
1.6	3	-91	-77	-97
	7	-72	-71	-79
	6	-63	-57	-71
6.1	3	-86	-77	-95
	7	-73	-69	-77
	6	-63	-53	-71
6.6	3	-87	-79	-93
	7	-75	-71	-81
	6	-52	-45	-71

5. Análisis de Resultados

Se posiciona el dispositivo en un punto del mapa (Figura 1) y se obtiene el vector de potencias de los AP visibles. Luego, con este vector patrón y con cada uno de los vectores de potencias almacenados (Media, Dupla intercuartílica, moda, promedio de dupla) se calcula la distancia euclidiana obteniendo las distancias a cada punto de coordenadas. Se determina la posición estimada como el menor valor que satisface la ecuación, es decir la menor distancia entre el conjunto de entrenamiento obtenido (base de datos *Fingerprint*) y el patrón de datos ingresado.

5.1 Posición (1,5):

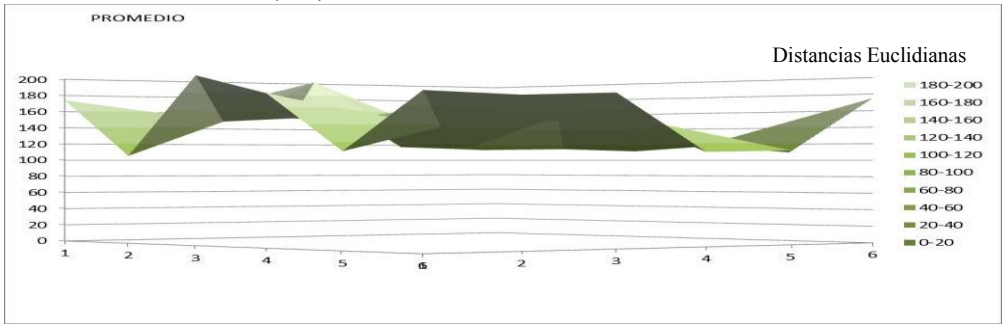


Figura 3. Gráfico de superficie PROMEDIOS posición (1,5)

Considerando la Figura 3, con el patrón: [0, 0, 0, -89, -61, -59, -75, -75, -67, -89, 0] y realizando el cálculo con los vectores promedios, la sección que minimiza la distancia es la posición de coordenadas (1,2), en este caso podemos observar que existe un error de 1.80 m, el cual considerando la tabla 1, puede ser originado por las fluctuaciones de las

señales de los AP, producida por la atenuación provocada por la pared adyacente al punto de muestreo, provocando disminución de RSSI e impidiendo la visibilidad de un AP 3.

5.2 Posición (5,3)

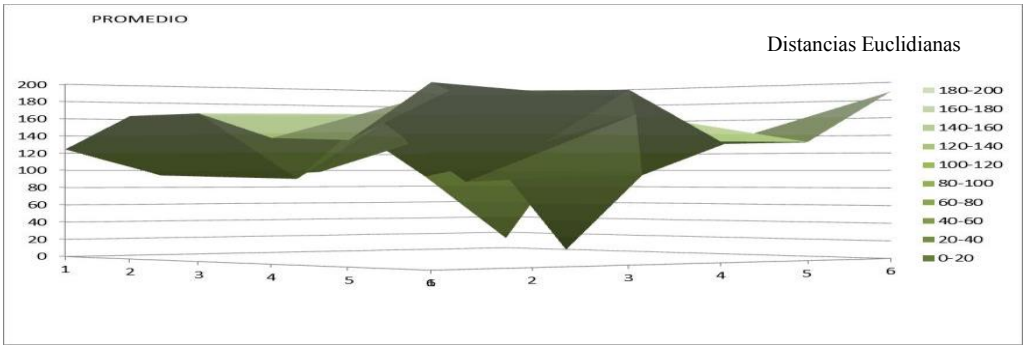


Figura 4. Gráfico de superficie PROMEDIOS posición (5,3)

La Figura 4 muestra el "*vector patrón*" de la posición (5,3): [0,0, -91, -45, 0, -63, -81, -83, -69, 0, -93]. La menor distancia obtenida por los promedios corresponde justamente a la posición 5.3, donde se observa un mínimo en el punto. Algunos vectores donde varía la señal de los AP más cercanos (3, 6 y 9); en un rango entre +- 2/4 dBm, la precisión, se reduce obteniendo la ubicación del DM en el punto (4,5) con un error de 1.7 m.

Observando el plano presentado en la Figura 1, se advierte que en ese entorno no existen paredes adyacentes al punto de captura, ver ejemplo (1,5) Figura 3.

5.3 Posición (4,5)

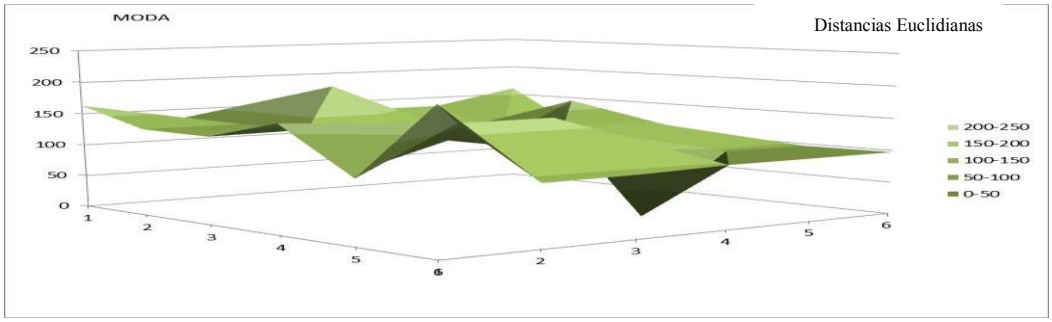


Figura 5. Gráfico de superficie MODA posición (4,5)

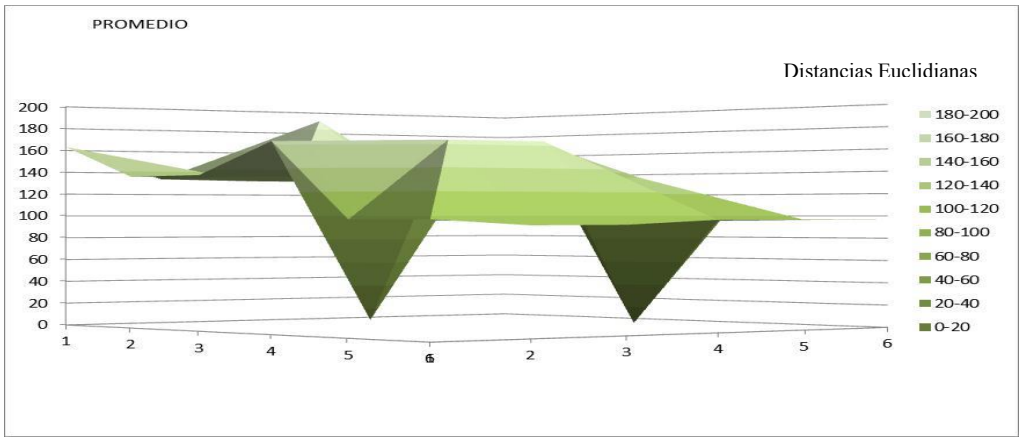


Figura 6. Gráfico de superficie PROMEDIO posición (4,5)

Las Figuras 5 y 6 muestran el patrón: [-95, 0, -93, -43, 0, -57, -77, -75, -65, -89, 0], correspondiente a la posición (4,5). Ambos gráficos coinciden en la posición, sin embargo, la Figura 5 obtenido por el cálculo de la distancia de los vectores moda almacenados en la base de datos, proporciona una mejor representación, originando un valor mínimo absoluto en el punto de posicionamiento. La Figura 6 calculada en base al promedio tiene otro resultado bastante cercano. En ambos ejemplos no existen obstáculos como paredes o ventanas adyacentes que provoquen un aumento de la absorción de la señal.

6. Conclusiones y Futuros Trabajos

Para analizar la tasa de error realizan 60 pruebas en cada punto del mapa (Figura 1), obteniendo el vector patrón para estimar la posición del DM. Del total de pruebas realizadas, se registró que el 70 % de las mismas obtenían las distancias presentadas en el gráfico de la Figura 7. Como se observa, en las coordenadas centrales del mapa es donde existe un menor margen de error, obteniendo como valor mínimo 1.2 m y como máximo 2.4 m. Al analizar los resultados, obtenidos en las secciones inferiores y superiores, se comprueba que el rango de errores varía entre 2.4 – 3.6 m.

De esta manera, se infiere que en las secciones donde existen determinados obstáculos adyacentes (paredes, ventanas, entre otros) los cuales causan el aumento de la absorción de la señal y del efecto multi-trayectoria es donde se presentan márgenes de errores mayores.

Se ha documentado una experiencia donde se ha logrado localizar un DM con un menor error trabajando con promedio y moda, sobre una base de datos *Fingerprint* con variantes.

Considerando toda la zona de análisis se logra posicionar con un error máximo promedio de 3.6 metros. En las zonas centrales, alejadas a unos 0.90 metros de las paredes, se logra posicionar con un error máximo de 1.7 metros.

La tecnología promete y se seguirá trabajando para reducir el error. Los próximos enfoques incluyen un método de votación para selección de la mejor técnica de forma automática, complementar el análisis considerando tiempo de vuelo de señal WIFI, pruebas en diversas oficinas, y pruebas en espacios abiertos.

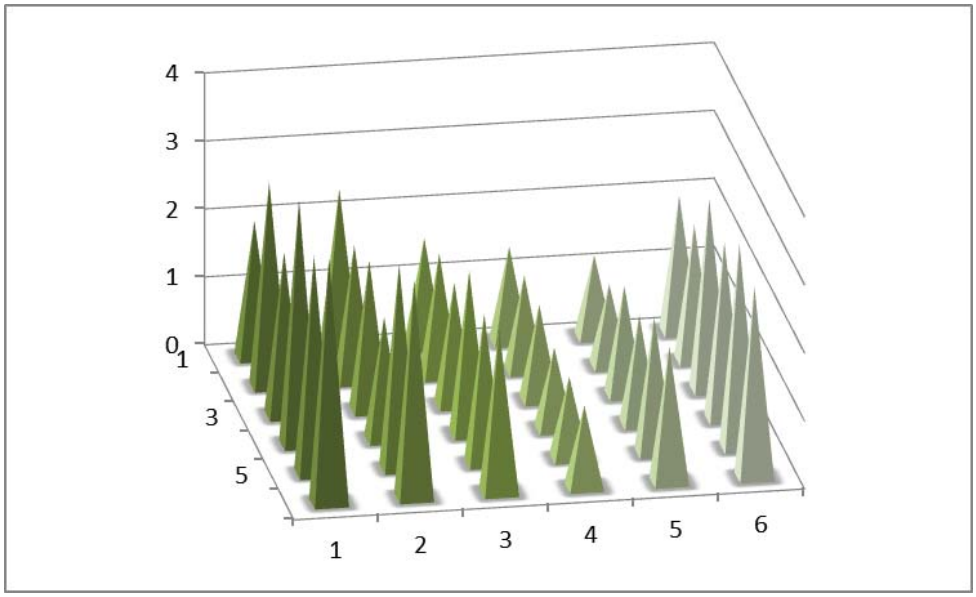


Figura 7. Errores obtenidos en el posicionamiento, donde se destaca cómo se incrementa el error cerca de las paredes

Referencias

[1] A. M. Ladd, K. E. Bekris, G. Marceau, A. Rudys, L. E. Kavraki, and D. S. Wallach, Robotics-based location sensing using wireless ethernet. ACM International Conference on Mobile Computing and Networking (MOBICOM'02), New York, 2002.

[2] G. M. Djuknic and R. E. Richton, Geolocation and assisted GPS, *IEEE Computer*. Vol. 34, Nro. 2, pp:123-125. 2001

[3**]** T. S. Rappaport, J. H. Reed, and B. D. Woerner, Position location using wireless communications on highways of the future, IEEE Communic. Vol. 34, Nro.10, pp: 33-41. , 1996.

[4] K. Pahlavan, X. Li, and J. P. Makela, Indoor geolocation science and technology, *IEEE Communications Magazine.*, Vol. 40, Nro. 2, pp: 112-118, 2002.

[5] P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF based user location and tracking system, *IEEE Infocom 2000*. Vol.2, Nro. 1, pp: 775-784. 2000.

[6] M. A. Youssef, A. Agrawala, and A. U. Shankar, WLAN location determination via clustering and probability distributions, in *Proc. IEEE International Conference on Pervasive Computing and Communications, Fort Worth,* 2003.
[7] A. Teuber, B. Eissfeller, WLAN indoor positioning based on Euclidean distances and fuzzy logic, *Proceedings of the 3rd Workshop on Positioning, Navigation and Communication (WPNC'06)*, Munich, Alemania, 2006.
[8] Ekahau, *Ekahau positioning engine 2.0; 802.11 based wireless LAN positioning system*, Ekahau Technology, Internal Report, www.eukahau.com, 2012.
[9] Roberto Battiti, Thang Lee Nhat, Alessandro Villani, Location-aware computing: a neural network model for determining location in wireless LANs, 2002.
[10] Pahlavan, K., & Krishnamurthy, P. Principles of Wireless Networks - A Unified Approach, Prentice Hall. 2002. ISBN-10: 0130930032
[11] Brachmann, F. A comparative analysis of standardized technologies for providing indoor geolocation functionality, Symposium on Computational Intelligence and Informatics (13th CINTI), 2012 IEEE, Hungary, Budapest 2012
[12] P. Enge and P. Misra, Special issue on gps: The global positioning system, Proceedings of the IEEE.Pp: 3-172. 1999.
[13] Marcelo Najnudel, Estudo de propagação em ambientes fechados para o planejamento de wlans, Universidad Católica de Rio de Janeiro. Tesis. 2004.
[14] J. Toloza, N. Acosta, A. de Giusti. An approach to determine magnitude and direction error in gps system. Asian Journal of computer science And Information Technology. Vol. 2, Nro. 9, Pp: 1-5. 2012.

Monitoreo remoto de sistemas y redes para la auditoria informática

María Elena Ciolli, Claudio Porchietto, Roberto Rossi, Juan Sapolski

Grupo de Investigación Instituto Universitario Aeronáutico, Córdoba, Argentina
{mciolli,porchietto,roberto.rossi}@gmail.com

Resumen. Esta ponencia presenta el resultado del análisis e implementación de herramientas para el control remoto del hardware y software de una red informática basado en la conceptualización GLPI (gestión libre del parque informático) y en la norma ISO 27002 dominio 7 (gestión de activos) sección 7.1(inventario de activos). Se realizó un estudio comparativo entre dos herramientas: OCS Inventory NG y Open Audit. Se tomaron como factores claves la identificación unívoca de hardware y el software del parque informático y asimismo se consideraron relevantes: el impacto en el tráfico de la red, las facilidades de las herramientas y la explotación de la base de datos resultante para su integración con otros sistemas de información.
Se pretende implementar un sistema de información automática de inventario que registre los cambios de la configuración de una red informática, aplicándose en primer término a la red interna del Instituto Universitario Aeronáutico que cuenta con un plantel de 1000 máquinas aproximadamente, repartidas entre dependencias del IUA central y centros de apoyo de Rosario y Buenos. Aires.

Palabras clave: OCS Inventory NG, Open Audit, GLPI, Auditoria, Monitoreo.

1 Introducción

Existen diversos estándares y prácticas [1] que definen cómo gestionar diferentes puntos de la función IT entre ellos :

- COBIT
- COSO
- ITIL
- ISO/IEC 27002
- FIPS PUB 200
- ISO/IEC TR 13335
- ISO/IEC 15408:2005
- TickIT
- TOGAF
- IT Baseline Protection Manual
- NIST 800-14

Fue seleccionada para esta investigación como base normativa la ISO/IEC 27002 [11],[12], por ser un estándar internacional en la cual los puntos de control son la clave para su implementación. En este proyecto se tomó de la misma el dominio 7, Gestión de activos, sección 7.1 ya que el mismo trata sobre Inventario de Activos y Directrices para su clasificación.

A los efectos de disponer de un estudio de campo que permita determinar el uso de aplicaciones GLPI en el entorno de las universidades tanto públicas como privadas de la ciudad de Córdoba Capital se ha realizado un relevamiento en distintas universidades, entre ellas la UNC y la UCC.

En este sentido se ha podido determinar que sólo en algunas áreas muy limitadas se utiliza software del tipo GLPI con fines de seguimiento de intervenciones sobre los equipos, como en el caso de soporte técnico, y no como gestión global de recursos informáticos, licencias de software o automatización del inventario.

En un diagrama como muestra la figura 1, pueden apreciarse los distintos roles que intervienen en una auditoría que realiza el control de activos informáticos:, a saber:

- Estaciones de trabajo.
- Auditor.
- Informe o reporte.
- Base de datos.
- Estación para el análisis de datos.
- Lista de procedimientos.

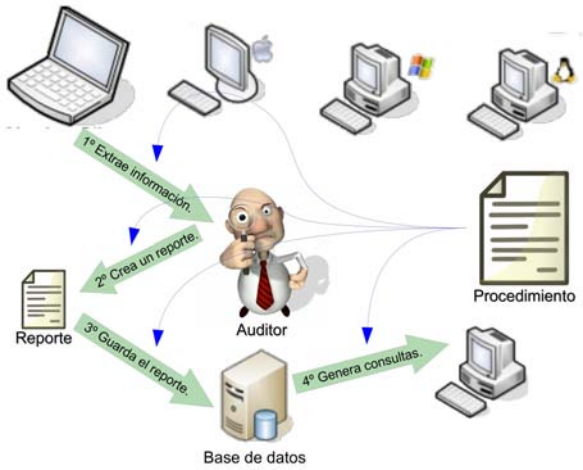


Figura 1. Auditoria estándar.

En la actualidad, en el organismo donde se realiza la investigación, el rol de auditor lo encarna una persona física apoyado por el software AIDA. El informe o reporte es transportado en un pendrive y la base de datos es una PC donde se guardan todos los informes. Todo esto se ejecuta en base a unos procedimientos internos estandarizados por normativas de la Fuerza Aérea Argentina, de la cual depende este instituto.

Como resulta evidente, es muy ardua la tarea de tener actualizada dicha base de datos, por lo que resulta imprescindible la investigación, desarrollo e implementación de un software que permita el control automático del parque informático de la institución y la generación y actualización de reportes mediante supervisión de la base de datos del mismo.

Se pretende, en síntesis, tener un control del inventario de la red informática tanto lógico como físico. Al realizarlo de manera autónoma, los períodos de actualización de la información resultan menores que cuando se realiza con un técnico que releva máquina por máquina en forma local y registra la información en una base de datos preexistente. Los beneficios más importantes son:

- Menor tiempo de actualización de la información.
- Disminución de la probabilidad de errores originados por el ingreso manual de los datos.
- Reducción de costos de mantenimiento.

2 Metodología

A la hora de implementar una solución al problema de la auditoría surgen distintas interrogantes, ¿qué Herramienta usar?, ¿cómo se implementa?, ¿qué datos se pueden extraer?, ¿qué datos son relevantes extraer?, entre otros.

Habiéndose analizado diferentes opciones para lograr este objetivo, se planteó un análisis de dos herramientas preseleccionadas de código abierto, a saber: OCS Inventory [2], [10] y Open Audit [9].

Con el fin de tener una primera aproximación al funcionamiento de las herramientas, este análisis fue llevado a cabo sobre un entorno de trabajo virtual. Posteriormente se realizó sobre una pequeña red LAN de arquitectura heterogénea

Nuestro esquema de funcionamiento está centrado en la auditoría de las máquinas que pertenecen a una red. En principio esta red está segmentada, con diferentes dominios, diferentes sistemas operativos, y diferentes usuarios. El primero de los interrogantes es ¿qué es necesario modificar o agregar a mi red para poder implementar el sistema de auditoría?

Luego surge la pregunta ¿cómo voy a enviar al auditor a cada estación de trabajo?.

Todas estas preguntas tienen un elemento en común que consiste en cómo factores externos a la herramienta afectan al despliegue de la misma [3]. De más esta decir que una herramienta de auditoría es netamente un sistema distribuido en toda la red.

Para responder estas preguntas es válida la utilización de un entorno de trabajo virtual.

El esquema de funcionamiento del sistema que se plantea se aproxima al que se muestra en la figura 2.

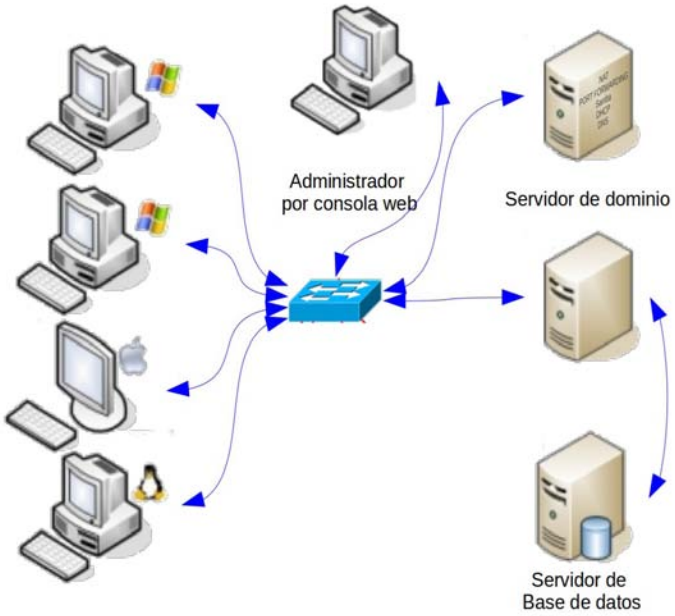


Figura 2. Estructura de la red virtual.

Esta estructura simula la red informática y se implementó en máquinas virtuales emuladas con Oracle VirtualBox.

¿qué datos se pueden extraer?

Al extraer datos tales como: usuarios, programas, configuraciones, etc, en general datos lógicos, la virtualización no presenta mayores inconvenientes, pero a la hora de extraer datos de los componentes físicos la misma no es suficiente.

De aquí surge la segunda etapa del proyecto, centrada en la fidelidad de los datos extraídos.

Nuestro nuevo entorno de trabajo es una pequeña red aislada, compuesta por 4 estaciones de trabajo todas de hardware diferente (distintos microprocesadores, placas madres, monitores, etc). Además cada estación de trabajo también contiene 4 sistemas operativos instalados. Si bien con solo 4 estaciones no se puede tener toda la diversidad de hardware que hay en una red de 1000 máquinas, esta configuración es una muestra representativa de un entorno real.

El esquema de funcionamiento del sistema que se plantea se aproxima al que se muestra en la figura 3.

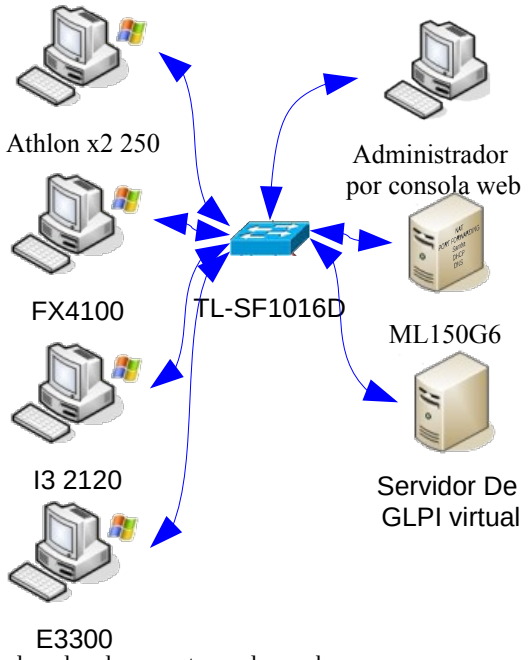


Figura 3. Topología de red real para entorno de pruebas.

Se puede apreciar en la figura 3 que el servidor de GLPI sigue siendo virtual. Esto no supone problemas a la hora de validar los datos ya que no es la máquina donde se alojan los servidores la que nos interesa auditar.

3 Resultados

De la primer etapa del proyecto surgen las guías de instalación y despliegue de las herramientas. Además se pudo estimar el impacto en la red para cada una de ellas. Se observó en un análisis de tráfico de red que el volumen de éste es directamente proporcional a la cantidad de máquinas. Esto nos permite estimar el tráfico total para la red de la institución. Con el fin de no congestionar a la red se programan los horarios y la velocidad de la auditoria. En la segunda etapa se realizó una comparación de la fidelidad de los datos extraídos, inspeccionando los informes de cada herramienta y verificando contra el hardware y software real de cada máquina.

Con todos estos informes se confeccionó la Tabla 1, donde se obtienen los siguientes indicadores, cuyos valores oscilan entre cero y cinco donde cinco es el máximo valor y cero el mínimo.

Tabla 1. Cuantificador numérico.

Herramienta, Atributo	*Valoracio de Importacia*	*OCS inventory Windows xp*		*OCS inventory Windows 7*		*OCS inventory ubuntu*	*Open Audit Windows xp*		*Open Audit Windows 7*		*Open Audit ubuntu*	
Inventario de Software												
Software de base con licencia -Sistema Operativo	5	5	5	4	4		0	5	5	4	4	0
Actualizaciones de Sistema Operativo	3	5	3	5	3	5	3	5	3	5	3	5	3
Software de aplicaciones con licencia	5	4	4	4	4		0	5	5	5	5	0
Antivirus	4	4	3,2	4	3,2		0	5	4	5	4	0
Software gratuito	4		0		0	5	4		0		0	5	4
Inventario de Hardware			0		0		0		0		0		0
Motherboard	5	2	2	2	2	2	2	4	4	4	4	4	4
Procesadores	5	4	4	4	4	4	4	5	5	5	5	5	5
Memoria	5	4	4	4	4	4	4	4	4	4	4	4	4
Almacenamiento fisico HDD	5	5	5	4	4	5	5	4	4	4	4	5	5
Almacenamiento fisico (CD, pen, etc)	5	4	4	4	4	4	4	4	4	4	4	5	5
Almacenamiento lógico	5	5	5	5	5	5	5	5	5	5	5	5	5
Video	5	3	3	3	3	3	3	3	3	3	3	3	3
Sonido	3	3	1,8	3	1,8	3	1,8	3	1,8	3	1,8	3	1,8
Red	5	5	5	5	5	5	5	5	5	5	5	5	5
BIOS	5	4	4	4	4	4	4	4	4	4	4	5	5
Monitor	4	5	4	5	4	1	0,8	5	4	5	4	5	4
Dispositivos de entrada.	3	4	2,4	4	2,4	2	1,2	4	2,4	4	2,4	4	2,4
Impresoras	4	4	3,2	4	3,2		0	2	1,6	2	1,6	4	3,2
Impacto en red			0		0		0		0		0		0
Volumen de tráfico en la auditoría	4	3	2,4	3	2,4		0	3	2,4	3	2,4	3	2,4
Volumen de tráfico en el despliegue	1	1	0,2	1	0,2		0	5	1	5	1	5	1
Facilidades			0		0		0		0		0		0
Desligue	3	5	3	3	1,8	5	3	5	3	5	3	5	3
TOTAL			68,2		65		49,8		71,2		70,2		65,8

Inventario de Software

- 5 corresponde a datos fidedignos y completos (nombre, versión, números de serie, licencia, etc,).
- 4 corresponde a datos fidedignos (nombre, versión, etc , pudiendo faltar algún número de serie o licencia pero auditando todo lo que tiene el sistema)
- 3 corresponde a datos parciales (ejemplo: No reconoce todo el software instalado o solo los nombres pero no las versiones)
- 2 corresponde a datos incompletos (Ejemplo: No detecta cierto software.)
- 1 corresponde a datos inciertos (Completa campos con nombres o números no significativos)

Inventario de Hardware

- 5 corresponde a datos fidedignos y completos (nombre, revisión, números de serie, etc,).
- 4 corresponde a datos fidedignos (nombre, modelo, pudiendo faltar algún número de serie, pero auditando todo lo que tiene el sistema)
- 3 corresponde a datos parciales (ejemplo: Reconoce cuanta memoria RAM tiene pero no el modelo.)
- 2 corresponde a datos incompletos (Ejemplo: No detenta un microprocesador, no detecta tarjetas de expansión.)
- 1 corresponde a datos inciertos (Completa campos con nombres o números no significativos)

Impacto de red

- 5 corresponde a volumen de tráfico excedente menor a la mitad al excedente promedio.
- 4 corresponde a volumen de tráfico excedente mayor a la mitad al excedente promedio.
- 3 corresponde a volumen de tráfico excedente cercano al excedente promedio.

• 2 corresponde a volumen de tráfico excedente menor al doble del excedente promedio.
• 1 corresponde a volumen de tráfico excedente mayor al doble del excedente promedio.

De la tabla 1 se puede concluir que Open Audit es la mejor elección.

¿cómo funciona Open Audit para auditar un dominio?

La figura 4 muestra un esquema general de auditoria de dominio con la herramienta Open Audit.

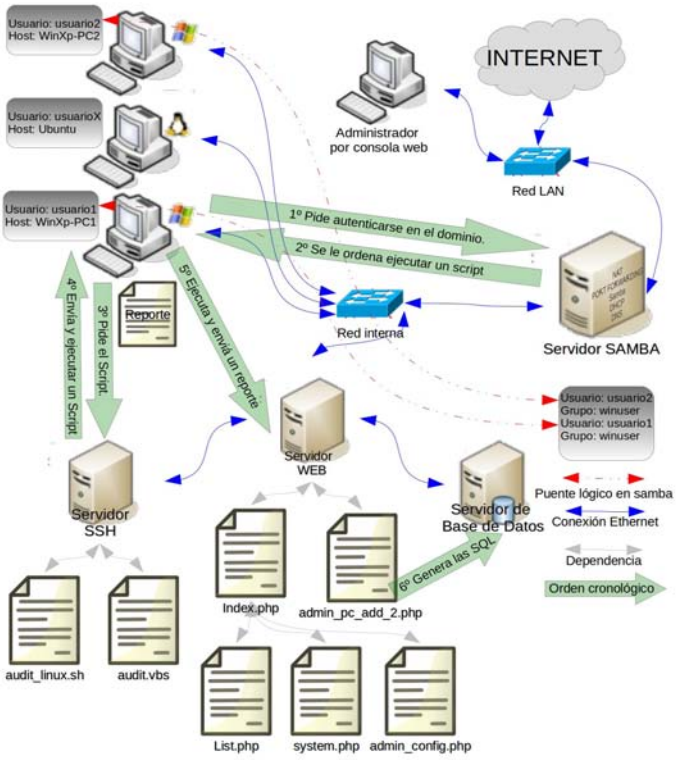


Figura 4. Auditoria de dominio.

Al iniciar sesión los usuarios del dominio se registran ante el PDC (controlador primario de dominio), que es implementado por el servidor Samba, que los instruye a ejecutar un Script de auditoria. Dependiendo del sistema operativo el Script es diferente.

El Script para Linux está compuesto de una serie de sentencias de consola cuya salida es analizada clasificada y segmentada por herramientas para procesar texto como awk.

En Windows se utiliza un Script semejante al de Linux que está codificado en Visual Basic y basado en instrucciones de WMI Service (Windows Management Instrumentation).

¿cómo funciona Open Audit para auditar máquinas fuera del dominio?

Un servidor con un método de Polling es el encargado de enviar el auditor. En la figura 5 se observa que aunque el segundo proceso de distribución parece más simple, no lo es, ya que es necesario cargar máquina por máquina en una lista con sus correspondientes nombres de usuario con privilegios de administrador.

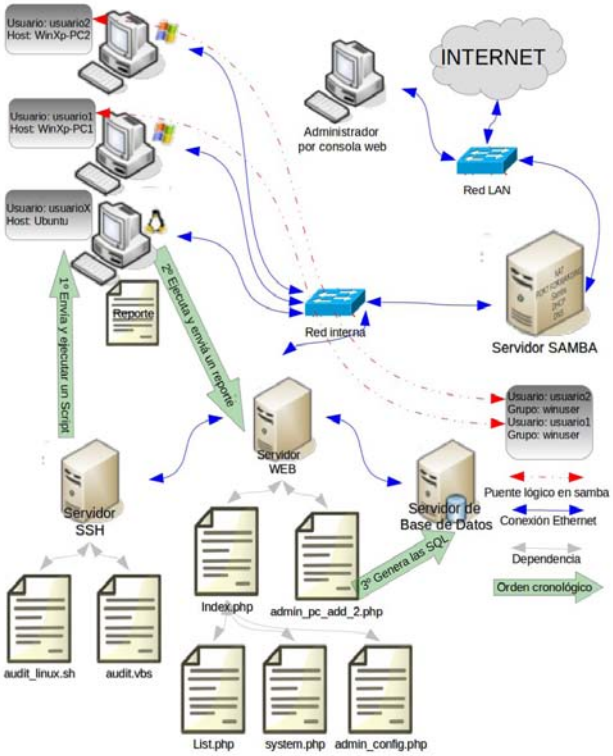


Figura 5. Auditoria fuera del dominio.

Todo esto conlleva a la necesidad de tener una gestión distribuida en la red y no concentrada.

¿cómo almacenan la información?

Ambos Scripts guardan en cadenas de texto la información extraída que contiene un identificador de cabecera y caracteres especiales como separador de campos. Estas cadenas son almacenadas temporalmente en un archivo de texto (reporte) cuyo nombre es el de la máquina auditada. Una vez realizadas todas las consultas, el archivo de reporte es enviado por html al servidor del Open Audit que se encarga de cargar los datos en el servidor de base de datos.

¿cómo se genera un reporte del estado general de la red?

En la figura 6 se aprecia cómo es el flujo de información a la hora de realizar una consulta. No es necesario que las máquinas auditadas estén en linea al momento de realizar la consulta. esto cumple con la disponibilidad de datos pedida por la ISO.

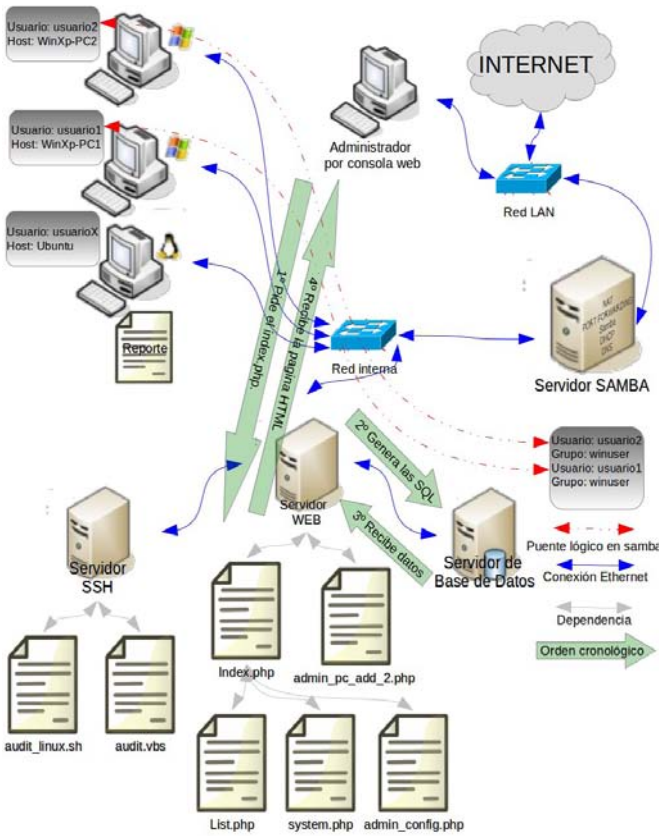


Figura 6. Consulta genérica.

En concordancia con el sistema actual, el rol del auditor es cambiado de una persona física a un Script, el reporte que se trasladaba y almacenaba manualmente ahora lo hace a través de la red LAN interna cuyo único requerimiento es que brinde conectividad entre las estaciones de trabajo y el servidor del Open Audit. Los procedimientos estandarizados están almacenados en el controlador de dominio que es quien va a decidir qué máquinas son auditadas y cuándo.

Esquema general

En la figura 7, se presenta un diagrama en bloques que muestra el esquema general que es necesario agregar a la red para implementar la herramienta.

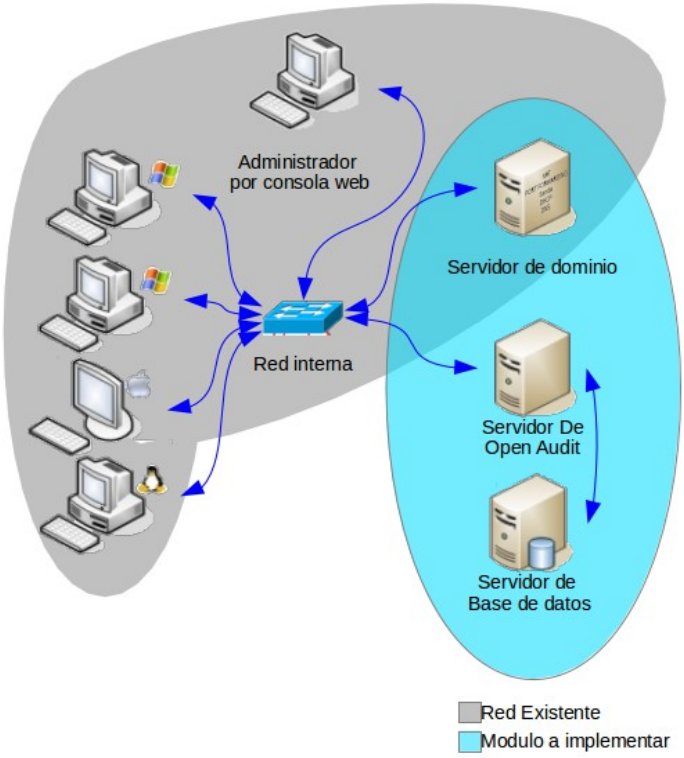


Figura 7. Modelo de implementación en red.

El área pintada de gris representa una red como la existente en el IUA, el área pintada de turquesa son los servidores que se incorporarán o modificarán.

Hay que destacar que hay un área compartida que es el servidor de dominio que al momento de implementar el sistema en la red real será necesario modificar su configuración. Por esto mismo es de vital importancia que estas configuraciones y modificaciones sean exhaustivamente probadas, a los efectos de evitar fallos en la red.

4 Conclusiones

Se generó un ambiente virtual donde se simuló un dominio real sobre el que se realizaron las pruebas de las herramientas de una forma controlada para poder medir bien sus facilidades.

El mismo está compuesto por máquinas virtuales emuladas con Oracle VirtualBox. Algunas de ellas actúan como servidores y otras como estaciones de trabajo, estas últimas fueron instaladas dentro de un mismo servidor, configurándolas en distintos ambientes operativos, para simular la situación real de la red del IUA, o de cualquier red informática con muchas estaciones de trabajo conectadas.

Además se generó un entorno de trabajo real que fue útil para verificar fidedignamente los datos que extraen las herramientas, el cual está compuesto por máquinas reales de hardware y software heterogéneo con el fin de tener una muestra más representativa del parque informático.

El éxito de estas pruebas originó que este logro técnico esté documentado y a disposición de los otros proyectos del Ministerio de Defensa en ejecución en la actualidad.

Las pruebas realizadas sobre el software demostraron que el mismo no es afectado por la topología de la red, ya que se parte de la presunción de que todas las máquinas tienen conectividad contra su servidor de dominios o la puerta de enlace a Internet, por lo que nos limitamos a simular solamente una subred: 10.0.0.x

Como se puede apreciar en la Tabla 1 la extracción de datos no cumple totalmente con lo requerido por la norma, por lo que es preciso continuar con el perfeccionamiento del código de Open Audit. Este es uno de lo motivos de que se elijan herramientas de trabajo de código fuente libre.

5 Trabajos futuros

Adaptación del frontend PHP que brinda la herramienta Open Audit con nuevas consultas SQL que faciliten la interacción con la información recolectada.

En la siguiente etapa se implementará una integración entre la base de datos de Open Audit y la base de datos de la institución a los fines de su convergencia a una única solución.

Referencias

1. http://auditoriasistemas.com/estandares-ti/
2. Barzan T. A. (2010). IT Inventory and Resource Management with OCS Inventory NG 1.02 , Ed. Packt Publishing.
3. Jackson C. (2010). Network Security Auditing. Ed. Cisco Press.
4. Fettig A. (2005). Twisted Network Programming Essentials. Ed. O'Reilly.

5. Philippe J. y Flatin M. (2002). Web Based Management of IP Network Systems. Ed. John Wiley & Sons.
6. McNab C. (2007). Network Security Assessment,
7. Echenique Garcia J. A.(2001). Auditoria en Informática. Ed. Compañía Editorial Continental.
8. Piattini V. M. y Del Peso N. E. (2008). Auditoria de Tecnologías y Sistemas de Información. Ed. Alfaomega Grupo Editor.
9. http://www.open-audit.org/
10. http://www.ocsinventory-ng.org/en/
11. http://www.iso27000.es/
12. http://www.17799.com/

DJBot: Administrando las salas de PC evitando la consola

Javier Díaz, Aldo Vizcaino, Alejandro Sabolansky y Einar Lanfranco

LINTI
Laboratorio de Investigación en Nuevas Tecnologías Informáticas,
calle 50 y 120, La Plata, Argentina
{javierd,asabolansky,avizcaino,einar}@linti.unlp.edu.ar
http://www.linti.unlp.edu.ar

Resumen La necesidad de optimizar las tareas de mantenimiento y uso remoto en las salas de PC de la Facultad de Informática de la UNLP dio lugar al desarrollo de DJBot: una aplicación web realizada en Python que permite ejecutar comandos en múltiples computadoras en un solo paso. En este documento se describen los motivos y el camino recorrido para llegar a la versión de la aplicación disponible hoy en día. DJBot se encuentra liberado como Software Libre para que cualquiera que desee pueda utilizarlo y contribuir en el proyecto. Desde su desarrollo evolucionó de ser una simple herramienta ejecutable en la línea de comandos a convertirse en una plataforma de administración centralizada que simplifica el mantenimiento, actualiza los equipos y permite ejecutar tareas programadas en la interfaz web.

Palabras clave: Botnet, Django, Fabric, Python, Redis

1. El contexto

La Facultad de Informática de la Universidad Nacional de La Plata (UNLP) actualmente cuenta con tres salas de PC, las cuales suman aproximadamente 80 equipos, que habitualmente se utilizan para dictar clases de las diferentes cátedras, realizar competencias, jornadas y otras actividades extracurriculares.

Todos los equipos disponibles cuentan con dos sistemas operativos instalados: Microsoft Windows 7 y la distribución de GNU/Linux que se desarrolla en la Facultad, Lihuen GNU/Linux[2], y queda a criterio de las cátedras y de los alumnos cuál utilizar.

Las tres salas se encuentran conectadas a la red troncal de la Facultad con acceso restringido desde el exterior, pero utilizando direccionamiento IPv4 público, lo que permite que los equipos sean identificados desde Internet.

El grupo de trabajo que conforman los autores de este documento, además de encontrarse a cargo de la administración y mantenimiento de las tres salas de PC, tiene acceso a los routers y demás dispositivos de interconexión dado que también es el grupo responsable del mantenimiento del enlace troncal de datos de la Facultad.

2. La problemática

Se pueden identificar dos cuestiones principales para resolver.

Por un lado, la problemática relacionada con las cuestiones rutinarias. La gran cantidad de actividades académicas que requieren el uso diario de las salas hace que la disponibilidad de espacio temporal para realizar tareas de mantenimiento en cada computadora sea casi nulo. Peor aún es el escenario que se presenta habitualmente donde los trabajos no se hacen sobre un equipo sino que se quieren realizar tareas en todos los equipos de la sala en un solo paso. Por ejemplo, un cambio de configuración o la instalación de un software requerido por algún usuario debe realizarse sobre todos los equipos de la sala.

Junto con la problemática planteada en relación a las tareas de mantenimiento, surgió la inquietud de cómo se podrían utilizar los equipos de la sala en los tiempos en que el equipamiento está ocioso para realizar alguna tarea específica como por ejemplo una prueba distribuida de carga a un servidor web.

Hasta el día de hoy, para realizar cualquier tipo de tareas era necesario ejecutar el software en forma manual por el operador en cada una de las máquinas involucradas mediante la interacción directa con cada uno de los equipos.

Para posibilitar y facilitar la realización en tiempo y forma de estas tareas se ha desarrollado DJBot, el proyecto que aquí se presenta.

3. La propuesta

Como respuesta a las problemáticas planteadas en la sección anterior se ha desarrollado una aplicación de administración simplificada y centralizada de terminales GNU/Linux la cual se controla mediante una interfaz web.

Todo ello fue desarrollado siguiendo el principio DRY (Dont Repeat Yourself) mediante la reutilización de componentes de software libre. A modo de resumen, se puede decir que DJBot fue realizado utilizando Python como lenguaje de programación, junto con diversos componentes y bibliotecas, como el framework Django para la interfaz de usuario, la biblioteca Fabric y el protocolo SSH, entre otros.

4. El camino

En un primer intento de solución, se utilizó Parallel SSH (PSSH)[1], una herramienta que permite entre otras cosas, realizar acciones mediante la ejecución de comandos y copiar archivos en diferentes computadoras en paralelo. PSSH se encuentra en los repositorios oficiales de la distribución Debian de GNU/Linux, lo que transitivamente hace que también esté disponible en Lihuen GNU/Linux y simplifica la puesta en producción del entorno.

Por su forma de funcionamiento, esta aplicación requiere que las direcciones IP de las máquinas a las cuales les solicita acciones se encuentren listadas en un archivo de texto. Con este archivo se ejecutan conexiones por SSH a cada una de las direcciones listadas. Por los mecanismos de protección propios de SSH[4] aparecieron restricciones:

- SSH funciona utilizando clave pública-privada. Los clientes –es decir, todas las máquinas cuya IP se encuentra en el archivo– deben contar con la clave pública del que se conectará para autorizarlo en el archivo /authorized_keys. Esto se solucionó propagando la clave pública a todos los equipos de todas las salas.
- SSH mantiene un archivo de confianza (/.ssh/known_hosts) donde guarda IP y clave pública de todos los pares con los con que dialoga en algún momento. Si ante un nuevo intento conexión el par no coincide, el servidor rechaza la conexión. Históricamente las máquinas de la Facultad estuvieron configuradas mediante un servicio de asignación dinámica de direcciones, de manera que la asignación de las direcciones IP se completaba en forma aleatoria. Con el servicio de SSH en funcionamiento, en cambio, el servicio de DHCP se modificó para que cada computadora pasara a tener una única IP fija y definitiva. Así, cuando se accedió por primera vez a las computadoras mediante SSH, se pudo generar el registro de identificación de cada máquina que asocia la computadora con su IP.

Una vez concluida esta etapa, se contaba con un software funcional que cubría en forma parcial las expectativas planteadas ya que permitía la administración remota de las salas. Sin embargo, la herramienta seguía sin cumplir las expectativas en su totalidad, tanto las originales como las que fueron surgiendo a medida que el desarrollo avanzaba.

En una segunda etapa del proyecto, surgió el desarrollo de una interfaz web como respuesta a la necesidad de que las tareas de mantenimiento y soporte pudieran ser realizadas desde cualquier lugar y por personal que no necesariamente

tenga acceso a la consola de administración. El hecho de que la implementación de PSSH exija la ejecución del intérprete de comandos BASH del sistema operativo GNU/Linux sumaba direccionamientos indirectos al proceso retrasando las tareas. Por esto, se comenzaron a estudiar alternativas y apareció la biblioteca Paramiko[5] –utilizada por PSSH– que no presenta la desventaja de los direccionamientos indirectos, pero solo permite establecer las conexiones de SSH de forma simple, lo cual dificulta el envio de órdenes a los clientes.

Por ello se descartó, y buscando una alternativa se optó por Fabric[6], una biblioteca que reúne las características más útiles de PSSH y que utiliza Paramiko para realizar las conexiones SSH.

En resumen, mediante Fabric se facilita la configuración de las tareas que se realizan en los equipos de las salas, facilitando la forma de indicar en qué terminales queremos ejecutar tareas, permitiendo utilizar distintos archivos de autenticación SSH y brindando la opción de elegir entre distintos usuarios, entre otras opciones.

Como última instancia restaba decidir cómo desarrollar la parte web; hoy es muy difícil pensar en desarrollar una aplicación web utilizando el lenguaje únicamente sin utilizar algún framework de desarrollo que acorte y simplifique el proceso mediante la provision de componentes genéricos como ser filtros de seguridad, plugins de autenticación, acceso a la bases de datos o mecanismos de templates para el desarrollo de las interfaces de usuario. Si bien existen numerosos frameworks disponibles para el desarrollo, el elegido en este caso fue Django[7], uno de los frameworks más difundidos y utilizados en el mundo del software libre; Seleccionado en este caso, fundamentalmente porque tiene la particularidad al igual que la biblioteca Fabric, de estar codificado en el lenguaje de programación Python[8].

Una característica particular de DJBot es su modo de funcionamiento. Las redes de zombis, más conocidas como "botnet" en el área de la seguridad informática, son redes formadas por equipos que realizan tareas automatizadas,donde hay un controlador principal, que generalmente a través de Internet que indica qué hacer a una gran cantidad de equipos "zombis" que siguen las ordenes sin preguntar el por qué. El diseño particularmente útil y simple que permite el framework de desarrollo Django, sumado al concepto básico de las "botnets", dieron lugar al nombre de esta aplicación: DJBot[11]. Como las partes que conforman su nombre lo indican, "DJ" hace mención a Django y "Bot" hace referencia al funcionamiento similar al de una "botnet". Así, DJBot es un sistema que integra manejo de conexiones SSH con una interfaz web.

5. El modelo de datos

El modelo de datos representado en la Figura 1 está compuesto por cuatro entidades: Aula, Computadora, Tarea y Configuraciones.

La entidad Aula está conformada por múltiples computadoras y presenta las siguientes características principales:

1. nombre del aula,

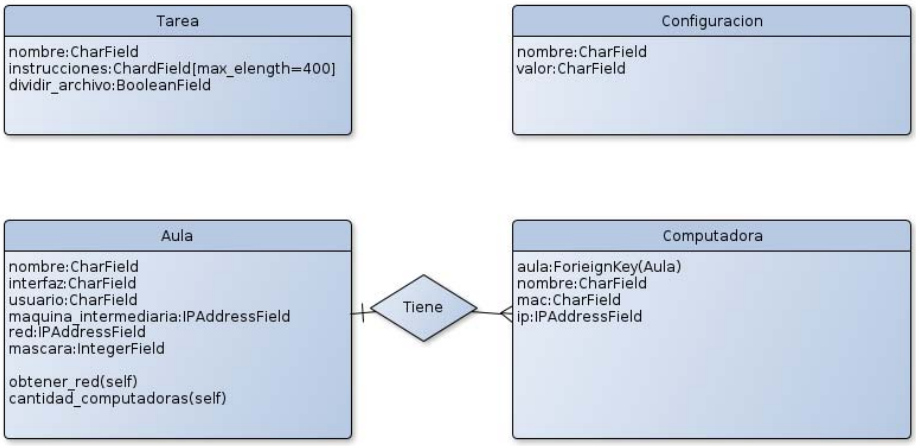


Figura 1. Clases del modelo de datos

2. usuario,
3. dirección IP de una máquina intermediaria,
4. nombre de la placa de red de la máquina intermediaria, y
5. dirección de red.

El nombre sirve para identificar el aula. La identificación mediante usuario permite realizar el inicio de sesión en cada computadora del aula. La dirección IP y el nombre de la interfaz de la máquina intermediaria, que por lo general es un router, permiten encender las computadoras de cada aula a través de la red (wake on LAN). La dirección de red, por su parte, se utiliza para asignar todas las máquinas encendidas dentro del rango de red indicado al aula específica.

La entidad Computadora incluye los valores:

1. nombre de la computadora,
2. dirección MAC,
3. dirección IP, y
4. aula a la que pertenece.

La entidad Tarea se define a través de:

1. nombre de la tarea,
2. conjunto de instrucciones, y
3. archivo opcional.

La lista de instrucciones completa la tarea en su totalidad. El archivo opcional permite compartir información entre todas las computadoras del aula, y se utiliza cuando resulta más sencillo que hacer uso de las instrucciones de comandos.

En la actualidad, la entidad Configuraciones se utiliza para definir valores predeterminados para las demás entidades. La clave SSH de las aulas, por ejemplo, está definida bajo la entidad Configuraciones. Sin embargo, en el futuro se

planea configurar claves SSH específicas para cada aula en particular. Así, la entidad Configuraciones quedaría disponible para cubrir cualquier necesidad de definición de valores predeterminados.

6. La optimización de DJBot

El desarrollo de DJBot, es decir, la integración del framework con las bibliotecas y con el modelo de datos que se describió anteriormente, se completó con el agregado de la interfaz web para lograr practicidad y flexibilidad en este trabajo.

La ejecución de tareas como la actualización del sistema y la instalación de aplicaciones nuevas, por ejemplo, implican tiempos de trabajo que al ser procesados desde la interfaz web de DJBot devolverían un error por tiempo de espera agotado. La utilización del buffer Redis[9] para almacenar las tareas solicitadas y los resultados devueltos evita, entonces, que el navegador quede como si estuviera cargando algo generándose retrasos innecesarios. Con la ayuda de la biblioteca RQWorker[10], Python se comunica con Redis y de esta manera las tareas se colocan en una cola de espera y son ejecutadas a su debido tiempo según el orden de solicitud. Esto independiza la interfaz web de DJBot para que el administrador pueda seguir interactuando con la misma sin tener demoras.

7. Los casos de uso

DJBot ya ha sido utilizado en la administración de las salas de PC de las facultad en diversas ocasiones, facilitando y automatizando las distintas tareas, entre las que podemos mencionar:

- Instalación de software solicitado por las cátedras para la realización de las actividades académicas.
- Actualización del sistema operativo ante actualizaciones de funcionalidad y seguridad de las distintas herramientas instaladas.
- Despliegue de una arquitectura para la realización de un test de stress sobre una aplicación web.

8. La liberación

DJBot está liberado como software libre bajo licencia GPL en su versión 3, una licencia que garantiza los principios del software libre permitiendo el libre uso, distribución y modificación del software a gusto de cualquiera que así lo desee.

Actualmente el código del proyecto puede encontrarse en los servidores de GitHub, uno de los repositorios de software libre más utilizados en la actualidad. Por como funciona GitHub, el lector puede descargarlo libremente desde allí; si es para usarlo puede hacerlo en forma anónima, pero en cambio si quiere generar algún aporte va a necesitar generarse un usuario en el sistema que le permita luego integrar su contribución al desarrollo.

Para acceder al desarrollo, solamente es necesario apuntar un navegador web a https://github.com/krahser/djbot.

9. El trabajo a futuro

Después de varios meses de trabajo y de tener la aplicación web funcionando, son varios los beneficios obtenidos y también son muchas las mejoras que están planificadas para ser aplicadas en el corto plazo.

Las cuestiones más importantes en las que se continuará trabajando involucran:

- Mejora de la interfaz gráfica para que resulte más simple de utilizar.
- Aplicación de mecanismos asincrónicos de notificación a través del uso de AJAX para agregar interacción a DJBot (ya se han realizado algunas pruebas con la biblioteca Dajaxice).
- Mejora de la seguridad del sistema mediante el uso de claves SSH independientes por sala.
- Agregado de un emulador de consola a la interfaz gráfica.
- Mejora del funcionamiento del buffer.
- Comprobación de la función de encendido de computadoras por red. Este ítem es necesario para evitar el tener que desplazarse hasta las salas para iniciar los equipos.
- Implementación de la función de descubrimiento de computadoras encendidas disponibles para trabajar en tiempo real.
- Carga automática de los valores de las computadoras que conforman una sala mediante una función de descubrimiento por red.

10. Las conclusiones

La disponibilidad limitada de las salas de PC de la Facultad de Informática motivó que desde el LINTI se desarrollara una aplicación web para poder realizar el trabajo de mantenimiento y administración en tiempo y forma sin interferir con la utilización académica habitual de las distintas salas de PC. Además de solucionar la problemática original, DJBot aportó soluciones colaterales, ya que no sólo permite ejecutar tareas y copiar archivos en más de una máquina al mismo tiempo, sino que ahora este trabajo no requiere de responsables expertos en la materia para poder hacerlo, ya que cualquiera puede invocar una tarea preprogramada desde la interfaz web. El acceso también es más simple; debido a que es una aplicación web, es posible acceder a la misma mediante un navegador web y así, sin más, tener acceso a todas las terminales salas de PC.

Practicidad y flexibilidad son apenas dos de los beneficios que brinda esta herramienta. Por otro lado, gracias a que ahora las computadoras de la Facultad están disponibles para ser utilizadas para realizar otras tareas en cualquier momento momento en que se encuentren ociosas y desde sitios remotos. También

se comenzaron a realizar tareas de investigación, como pruebas de estabilidad y carga distribuida de servicios web.

La práctica y experiencia de uso que se ha tenido hasta el momento son aspectos motivadores para continuar mejorando DJBot y seguir enfrentando los desafíos diarios con creatividad y precisión. DJBot es tan solo un ejemplo del amplio abanico de soluciones que están a la espera de ser creadas.

Referencias

1. Parallel SSH, http://code.google.com/p/parallel-ssh/
2. GNU/Linux Lihuen , http://www.lihuen.linti.unlp.edu.ar
3. Dinamic Host Configuration Protocol, http://linux.die.net/man/5/dhcpd.conf
4. SSH Known Hosts, http://www.linuxmanpages.com/man1/ssh.1.php
5. Paramiko, https://github.com/paramiko/paramiko
6. Fabric, http://docs.fabfile.org/en/1.6/
7. Django, versión 1.4.5, https://www.djangoproject.com/
8. Python, versión 2.7, http://www.python.org/
9. Redis, http://redis.io/
10. Django rqworker, https://github.com/ui/django-rq/
11. DJBot, https://github.com/krahser/djbot

V WORKSHOP INNOVACIÓN EN SISTEMAS DE SOFTWARE

- WISS -

CACIC2013

V WORKSHOP INNOVACIÓN EN SISTEMAS DE SOFTWARE

- WISS -

ID	Trabajo	Autores
5673	**Gestión del Conocimiento: Un enfoque aplicado en la Administración Pública**	Sebastián Pardo (HTC-GBA), Juan Enrique Coronel (HTC-GBA), Rodolfo Bertone (UNLP), Pablo Thomas (UNLP)
5735	**Aplicación de los condicionales DHD en la herramienta del foro de Sakai**	Alejandro Sartorio (CIFASIS), Marcelo Vaquero (UAI), Guillermo Rodriguez (CIFASIS), Daniel Tedini (UAI)
5736	**Framework para Interfase Cerebro Computadora implementado para el control de artefactos en el contexto de la domótica**	Jorge Ierache (UM), Gustavo Pereira (UM), Juan Iribarren (UM), Enrique Calot (UBA)
5761	**Una aplicación de la Wikimedia Semántica**	Marcela Vegetti (UTN), Horacio Leone (UTN)
5772	**Applying EDON Methodology and SBVR2OWL Mappings for Building an Ontology-Aware Software**	Cecilia Gaspoz (UTN-FRSF), Valeria Bertossi (UTN-FRSF), Emiliano Reynares (UTNFRSF), María Laura Caliusco (UTN-FRSF)
5778	**Propuesta de tecnología móvil para la administración de información vinculada a la gestión de espacios áulicos**	Martín S. Martínez (UNNE), Sonia Mariño (UNNE), Pedro Luis Alfonzo (UNNE), María V. Godoy (UNNE)
5827	**Desarrollo de aplicaciones colaborativas para Cloud Computing**	María Antonia Murazzo (UNSJ), Nelson R. Rodriguez (UNSJ), Daniela A. Villafañe (UNSJ), Daniel Gallardo (UNSJ)
5794	**"ASLX": Semantic Analyzer for programming languages based on XML**	Yanel Buffa (UNRC), Franco Gaston Pellegrini (UNRC)

V WORKSHOP INNOVACIÓN EN SISTEMAS DE SOFTWARE

- WISS -

ID	Trabajo	Autores
5808	**Sistema Guía para Personas con Deficiencia Visual**	Pablo Richard (UNCa), Daniel Richard (UNCa), Marcos Aranda (UNCa)
5812	**Automatización en la Captura de Datos para el Modelado de Flujo Vehicular**	Julio Monetti (UTN-FRM), Mariana Brachetta (UTN-FRM), Oscar León (UTN-FRM)
5887	**Planeamiento Estratégico para Compartir Información en la Administración Pública**	Ignacio Marcovecchio (UNS), Elsa Estevez (UNN), Pablo Rubén Fillottrani (UNS)
5824	**Ciber-adicciones: Estudio del Comportamiento Poblacional por Simulación**	L. Montesano (UTN-FRBA), Maria Florencia Pollo Cattaneo (UTN-FRBA), Ramon Garcia Martinez (UNLA)
5846	**Software e innovación: desarrollando productos con hardware y software flexible**	Daniel Díaz (UNSJ), Sandra Oviedo (UNSJ), Leandro Muñoz (UNSJ), Francisco Ibáñez (UNSJ)
5848	**Un Marco de Trabajo para la Integración de Arquitecturas de Software con Metodologías Ágiles de Desarrollo**	Luis Vivas (UNRN), Mauro Cambarieri (UNRN), Marcelo Petroff (UNRN), Horacio Muñoz Abbate (UNRN), Nicolás García Martinez (UNRN)
5888	**Herramienta de gestión de trazabilidad de requerimientos en proyectos de software←/td→**	Alfredo Villafañe (UNNE), María de los Ángeles Ferraro (UNNE), Yanina Medina (UNNE), Cristina Greiner (UNNE), Gladys N. Dapozo (UNNE), Marcelo Estayno (UNLZ)

Gestión del Conocimiento: Un enfoque aplicado en la Administración Pública

Sebastián Pardo[1], Juan Enrique Coronel[1], Rodolfo Bertone[2], Pablo Thomas[2]
[1] Honorable Tribunal de Cuentas de la Provincia de Buenos Aires – Argentina
[2] Instituto de Investigación en Informática LIDI - Facultad de Informática
Universidad Nacional de La Plata - Argentina

{spardo, jcoronel} @htc.gba.gov.ar
{pbertone, pthomas}@lidi.info.unlp.edu.ar

Abstract. La Gestión del Conocimiento abarca al conjunto de actividades realizadas con el fin de utilizar, compartir y desarrollar el conocimiento de una organización y de los individuos que en ella trabajan, encaminándolos a la mejor consecución de sus objetivos. En este trabajo se presenta un enfoque de estudio de Gestión del Conocimiento y se efectúa, a su vez, un análisis de la Gestión del Conocimiento aplicada en diferentes ámbitos de la Administración Pública. A modo práctico, se ha desarrollado una herramienta de software que promueve la Gestión del Conocimiento, la cual se implantó de manera concreta en un Organismo Público, el Honorable Tribunal de Cuentas de la Provincia de Buenos Aires (HTC). Aplicar las teorías de Gestión del Conocimiento permitir visualizar, compartir y utilizar de diversas maneras los recursos no tangibles existentes, en pos del progreso y modernización de las Organizaciones Públicas.

Keywords: Gestión del Conocimiento - Unidad de Información - Honorable Tribunal Cuentas Provincia Buenos Aires - Gobierno Electrónico.

1 Introducción

Conocimiento es aquello que permite tomar decisiones y actuar [1]. Puede definirse la Gestión del Conocimiento como el conjunto de actividades realizadas con el fin de utilizar, compartir y desarrollar el conocimiento de una organización y de los individuos que en ella trabajan, encaminándolos a la mejor consecución de sus objetivos [2].

La Gestión del Conocimiento genera recursos para las organizaciones, el denominado capital intelectual, como elemento intangible y perdurable para una gestión eficiente y sostenible en el tiempo. Mediante la Gestión del Conocimiento las organizaciones favorecen que el individuo se desarrolle en su trabajo aportando ideas, y al mismo tiempo se evita la fuga de conocimiento cuando las personas abandonan la organización. Este concepto de retención del conocimiento tácito o inconsciente representa una innovación en el campo de la administración y se hace imprescindible en la actualidad.

Es necesario enfocarse en la necesidad de Estados inteligentes y eficientes: relacionar y unificar criterios, descubriendo que en la administración pública hay consumidores de servicios (ciudadanos), competencias y proveedores, como así también regionalizaciones y necesidades de integración entre todos esos componentes.

El objetivo propuesto en este trabajo es el desarrollo de una plataforma informática colaborativa, que permita la gestión sistematizada de Unidades de Información en el Honorable Tribunal de Cuentas de la Provincia de Buenos Aires.

La innovación de la plataforma consistirá en incorporar en el diseño los conceptos de participación y Gestión del Conocimiento a través de la interrelación de las Unidades de Información, como así también permitir al usuario la ponderación de cada Unidad de Información, la generación de comentarios relacionados con la unidad en cuestión, enlaces a sitios web y referencias bibliográficas.

A continuación se presenta un enfoque conceptual de la Gestión del Conocimiento, posteriormente se describe la Gestión del Conocimiento en la administración pública, y luego se presenta el sistema de software desarrollado para el HTC. Finalmente se detallan resultados obtenidos, conclusiones y trabajo futuro.

2 Gestión del Conocimiento: un enfoque conceptual

2.1 El Conocimiento y las Organizaciones

Davenport y Prusak establecieron una distinción entre tres conceptos: dato, información y conocimiento [6]. Los datos son la mínima unidad semántica, y se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones.. La información se puede definir como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre. El conocimiento es una mezcla de experiencia, valores, información y know-how que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción. Se origina y aplica en la mente de los conocedores [3].

Desde la perspectiva de la Gestión del Conocimiento, uno de los aspectos de la epistemología de mayor relevancia es el del proceso de generación y adquisición de conocimiento.

Davenport y Prusak definen al conocimiento como una mezcla fluida de la experiencia acumulada, los valores, la información contextualizada y la intuición del experto que crea un marco de referencia para la evaluación, y la incorporación de nuevos aprendizajes y de información. Dentro de las organizaciones, el conocimiento se encuentra inmerso en los repositorios, pero también en los procesos organizacionales de rutina, en sus prácticas y en sus normas [4].

Nonaka y Takeuchi generaron un desarrollo conceptual donde el conocimiento se crea realmente cuando los tipos de conocimiento se convierten entre sí y de uno a otro, a través de los niveles organizacionales, comenzando en el individuo y

ascendiendo al ámbito grupal, organizacional e ínter organizacional, creándose una espiral que produce la innovación no sólo en productos y tecnologías, sino también en procesos y estrategias organizativas [5]. Este enfoque configura el pensamiento dominante sobre el tema en la actualidad.

2.3 Metas, Objetivos y Pilares para la Gestión del Conocimiento

La meta primaria de la Gestión del Conocimiento se define como la mejora de las prestaciones organizativas por la captación de los individuos para obtener, compartir y aplicar conocimiento colectivo para tomar decisiones óptimas en tiempo real, entendiéndose este último como el tiempo disponible para tomar la decisión y ejecutar la acción que afectará materialmente el resultado.

En cuanto a los objetivos, pueden enumerarse los siguientes:

1. Hacer que las instituciones en general y empresas en particular actúen tan inteligentemente como sea posible para asegurar su viabilidad y éxito global.
2. En otro caso, darse cuenta del mejor valor de sus activos de conocimiento.

Para alcanzar estas metas, las organizaciones construyen, transforman, organizan, despliegan y suman efectivamente activos de conocimiento [6].

Un sistema de Gestión del Conocimiento debe conjugar tres pilares fundamentales para la gestión:

- El personal y la cultura.
- La gestión institucional.
- La tecnología. (portales, Groupware, herramientas de comunicación electrónica, almacenes de datos y minería de datos, y la infraestructura de soporte).

3 Gestión del Conocimiento en la administración pública

Resulta útil distinguir los conceptos de sociedad de la información y del conocimiento. La sociedad de la información hace referencia a la creciente capacidad tecnológica para almacenar más información y hacerla circular cada vez más rápidamente y con mayor capacidad de difusión. La sociedad del conocimiento se refiere a la apropiación crítica y selectiva de la información protagonizada por ciudadanos que saben cómo aprovechar la información [7]. Sobre la base de estas sociedades se cimientan las instituciones públicas.

Puede afirmarse que las instituciones públicas son grandes productores y consumidores de conocimiento. Al contrario de lo que ocurre con la empresa privada, la administración no tiene que preocuparse de la rentabilidad sino que debe prestar especial atención a dos aspectos esenciales [1]:

- Ser altamente eficiente en recaudar y gastar adecuadamente los recursos.
- Mejorar la calidad de vida de sus ciudadanos mediante los servicios especializados que prestan.

Sin embargo, para lograr dichos objetivos, las instituciones públicas tienen a su vez graves problemas para abordar otros dos aspectos:

1. Precisar con detalle los resultados que prometen obtener, y más en concreto, los indicadores de gestión que dan cuenta de su cumplimiento.
2. Determinar cuál es el conocimiento crítico que mayor impacto tiene en el logro de dichos resultados.

Además existen los factores políticos tales como:

- Problemas internos: autoritarismo, corrupción, desigualdades sociales, ineficiencia, insuficiente organización de las instituciones públicas, entre otros.
- Problemas externos: centralización institucional de los Gobiernos, leyes y regulaciones que limitan los municipios y estados, cuestiones político-partidarias relacionadas a aspectos presupuestarios, entre otros.

Las organizaciones públicas son básicamente organizaciones del conocimiento y para cumplir con su rol, la materia prima con la que trabajan es básicamente información, y el servicio que entregan al cliente es conocimiento depurado. Puede afirmarse que en la actualidad la inmensa mayoría de organizaciones carecen de una estrategia definida para gestionar su activo primordial [8].

4 Sistema de Software para la Gestión del Conocimiento en el ámbito de la Administración Pública de la Pcia. de Bs. As.

4.1 Contexto de Aplicación

El Honorable Tribunal de Cuentas de la Provincia de Buenos Aires (HTC) es un Organismo Constitucional. Sus atribuciones, tipificadas en la ley provincial N° 10869 (Orgánica del Tribunal de Cuentas), son [9]:

1) Examinar las cuentas de percepción e inversión de las rentas públicas, tanto provinciales como municipales, aprobarlas o desaprobarlas y en este último caso, indicar el funcionario o funcionarios responsables, como así también el monto y la causa de los alcances respectivos.

2) Inspeccionar las oficinas públicas provinciales o municipales que administren fondos públicos, y tomar las medidas necesarias para prevenir cualquier irregularidad en la forma y con arreglo al procedimiento que determine la Ley.

4.2 Génesis del Proyecto

El HTC contaba con distintos sistemas descentralizados para el tratamiento de Doctrinas, Jurisprudencias y Fallos, temas específicos que son atendidos por las Secretarías Jurídica, de Consultas y General respectivamente, dependientes de la Presidencia del Organismo. Estos sistemas se encontraban implementados en lenguajes de programación obsoletos, con motores de bases de datos de distintas tecnologías, cuyos ciclos de vida tecnológicos ya habían finalizado. Por ejemplo, existían tres implementaciones desarrolladas en Visual Basic 6, cuyo soporte finalizó

en 2005 y fue extendido hasta 2008. En cuanto a las bases de datos, se presentaban implementaciones en Microsoft Sql Server 7 y Microsoft Access.

Estas implementaciones requerían un constante mantenimiento correctivo, como así también se presentaban nuevos requerimientos funcionales acordes a las necesidades tecnológicas actuales, como por ejemplo la centralización y catalogación de publicaciones, agilidad en la carga y estadísticas.

Los requerimientos e inquietudes de los usuarios comenzaron a surgir con fuerza, sumándose a la necesidad de cambio propia de la dinámica de un área de sistemas. Esto motivó un relevamiento inicial y una tormenta de ideas, contemplándose dos alternativas de solución:

1. Solución Convencional: generar un nuevo sistema o módulo para cada secretaría, según prioridades a analizar. Esta solución continuaría con el paradigma descentralizado de la información del organismo.
2. Solución basada en Gestión del Conocimiento: generar un sistema que, a partir de relevar específicamente los requerimientos de las secretarías, presente una solución integrada para una gestión sistematizada de todas las publicaciones, basándose en los principios de la Gestión del Conocimiento. Esta solución requiere atender un análisis integral de las secretarías, su funcionamiento, flujo de información, necesidades y propuestas, así como también un relevamiento y análisis de fortalezas, oportunidades, debilidades y amenazas de los sistemas ya implementados, con el fin de receptar y aprovechar las experiencias anteriores. Técnicamente, implementar un sistema de este tipo demanda un grado de demora mayor para la puesta en producción, debido a las dificultades naturalmente derivadas de la unificación de criterios y requisitos, mayores tiempos de análisis, y dificultades para la implementación de sistemas genéricos.

En ambos casos se debe migrar toda la información posible desde los sistemas obsoletos hacia las nuevas implementaciones.

A partir de un análisis final, se resolvió implementar la segunda solución, atendiéndose los costos que demanda una solución integral.

4.3 Summun: solución integral y basada en Gestión del Conocimiento

Summun, el producto de software desarrollado, se define como una herramienta de construcción participativa, transversal, escalable y fundacional de una estrategia de inteligencia y aprendizaje organizacional [10]. Para el desarrollo se utilizó tecnología "Symfony 2", como un framework basado en software libre y diseñado para optimizar el desarrollo de las aplicaciones web basado en el patrón MVC (Modelo, Vista, Controlador).

Para la base de datos se definió un modelo de datos orientado a objetos, y se utilizó el DBMS MySql, a través del ORM de Doctrine, el cual permite asociar objetos a una base de datos relacional [12].

4.4 Fases de Desarrollo

Se definieron 3 fases de desarrollo:

Fase I: "Gestión Individual de Unidades de Información"

Esta fase define lo que se denomina "Unidad de Información" (UI) que consiste en el componente básico o primario de Información a gestionar por el sistema. La gestión abarca las Altas, Bajas, Modificaciones, Búsquedas Simples y Avanzadas, Manejo de Estados(Borrador, Cargado, Visado, Acceso Interno, Acceso Público), Palabras Claves, Sectorización de Usuarios (carga y visado) y Trazabilidad de Doctrinas, Jurisprudencias, Fallos Judiciales, Fallos del Tribunal de Cuentas y Normativas. Las Unidades con estado "´Acceso Público" deben ser posibles de ser accedidas por el público en general a través del sitio web del organismo.

Fase II: "Gestión Participativa y Construcción de Conocimiento"

En esta fase se hacen explícitos los conceptos de Gestión del Conocimiento. Se incluyen elementos como [10]:

- Relaciones: permiten vincular unidades de diversos modos.
- Enlaces Externos: permiten vincular una unidad con uno o más hipervínculos.
- Referencias bibliográficas: consisten en citar de la manera más detallada posible referencias en libros, manuales, tratados, cartillas o cualquier otro tipo de compendio físico que se halle disponible en el organismo.
- Comentarios: se trata de un campo texto, en el que cualquier usuario registrado puede dejar comentarios, que permitan acrecentar el valor de la unidad.

Fase III: "Gestión Integrada"

La tercera fase eleva el nivel de complejidad y planea generar "digestos" que integren diferentes UI para fines específicos. Pueden darse implementaciones vinculadas con la gestión de calidad del Honorable Tribunal de Cuentas.

4.5 Implementación de Summun

En la página de inicio, presentada en la figura 1, se visualiza un panel de control donde se muestran datos estadísticos y, asimismo, información concerniente a las UI más recientemente utilizadas. La "Unidad de Información" consiste en el componente básico o primario de información a gestionar por el sistema. La gestión de las UI representa el alimento de datos de este software, la fuente de información que luego proveerá a los futuros sistemas para la toma de decisiones, soportados en Summun.

El Panel de Control introduce una innovación en el campo de los Sistemas informáticos del HTC. Se incorpora este concepto al desarrollo del software, el cual consiste en concentrar en una sola página y en tiempo real todos los aspectos que se consideran importantes en la gestión. Se presenta como una herramienta ejecutiva y de gestión que muestra información importante de manera centralizada.

Puede encontrase elementos tales como: Últimas unidades cargadas, Gráfico comparativo de cantidades, Ranking de Unidades más Visitadas, Unidades Creadas, Ranking de Usuarios Creadores y Listado de Marcadores.

Summun presenta en su menú de UI un listado con las distintas unidades a gestionar, a saber: Jurisprudencia, Doctrina de Consultas, Doctrina de Jurídicas, Fallos del HTC y Normativas.

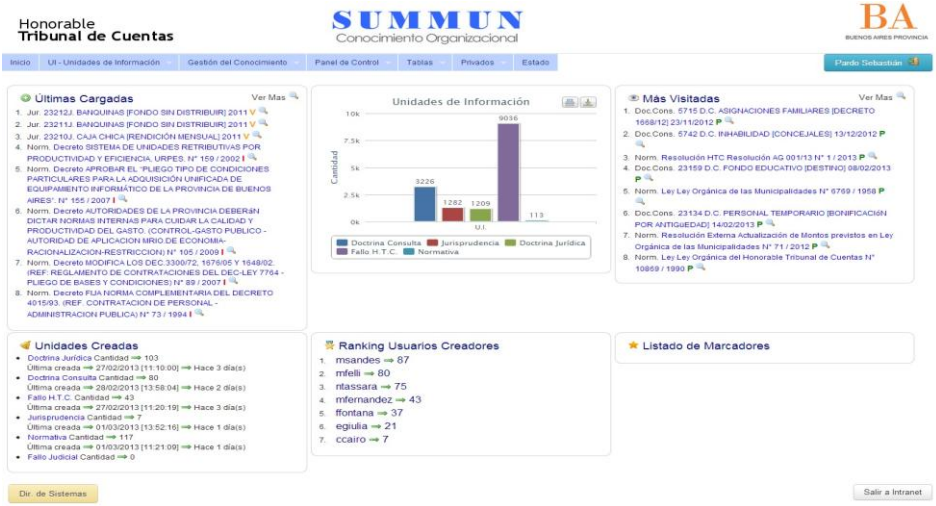


Figura 1: Pantalla inicial de Summun

El estilo de diseño de todos los tipos de UI respeta el mismo concepto, y es el de presentar una pantalla con todas las unidades de información correspondiente al tipo de unidad en cuestión, y un conjunto de funcionalidades representadas en la figura 2:

Figura 2: Módulo de ABM de Doctrina de Consultas

Para el ejemplo, se cuenta con una página principal organizada de modo tal de mostrar en forma de grilla las unidades de información disponibles y sus características, como así también el agrupamiento de un conjunto de acciones para cada unidad. Además se dispone de accesos para efectuar búsquedas.

Summun incorpora en su diseño de conceptos de "Participación" y "Gestión del Conocimiento". Esto último se logra a partir de interrelacionar las UI y permitir operatoria vinculada con la definición de Gestión del Conocimiento tal como lo son las ponderaciones, comentarios, relaciones y referencias.

Esta combinación de carga de datos, relaciones, comentarios y referencias, logran un círculo virtuoso para la Gestión del Conocimiento del Honorable Tribunal de Cuentas que se soporta en el sistema informático presentado, respetando la definición. Cada Unidad de Información posee características propias relacionadas con la Gestión del Conocimiento.

La funcionalidad de Gestión del Conocimiento permite acceder a una interfaz de consulta de todo lo inherente a la Gestión del Conocimiento vinculado a la Unidad. A modo de integración, se presenta en las figuras 3 y 4 la funcionalidad de Gestión del Conocimiento asociada a una unidad.

Figura 3: Enlaces y Relaciones de una UI

Figura 4: Comentarios, Referencias y Ponderación de una UI

En las figuras pueden visualizarse los enlaces, relaciones, comentarios y referencias bibliográficas, aspectos ligados al menú de Gestión del Conocimiento. Este concepto está implementado dentro del contexto de la etapa 2 de Summun. Se subdivide en cuatro funcionalidades, cada una de las cuales construye la idea de relaciones, comentarios, enlaces y referencias, como principios claves para la Gestión del Conocimiento en el Honorable Tribunal de Cuentas.

Además se permite ponderar a la Unidad de Información según su utilidad, confiabilidad y completitud. La ponderación es un concepto inspirado en las evaluaciones de Wikipedia, y permite a futuro generar un panel de control donde puedan incluirse rankings con las "mejores" unidades de información de acuerdo a cada criterio.

5 Resultados Obtenidos

A partir del desarrollo de un módulo especial, se migraron unidades documentales y archivos físicos con una efectividad cercana al 100%; exactamente 14.504 Unidades de Información obtenidas de sistemas de gestión de bases de datos PostgreSQL, SQL Server y MySql. En 6 meses de utilización del sistema se cargaron aproximandamente 800 Unidades, lo que resalta la importancia de la migración en cuanto a volumen de datos.

Al 3 de Julio de 2013, se registraron 14756 visitas públicas y 1087 de uso interno, lo que refleja el alto grado de participación por parte de la sociedad.

El HTC posee la certificación ISO 9001:2008, la cual contempla la implementación de "No Conformidades" como instrumento para que los usuarios puedan evidenciar incumplimientos de requisitos. Al 3 de Julio de 2013 no existen No Conformidades vinculadas al Sistema Summun.

Asimismo la Dirección de Sistemas posee un software de soporte para que los usuarios envíen sus problemas técnicos, necesidades, requerimientos u opiniones. A partir del gestor IT y los registros de asistencia telefónica al usuario, en la figura 5 se presenta las solicitudes y su distribución, al 12 de Marzo del 2012:

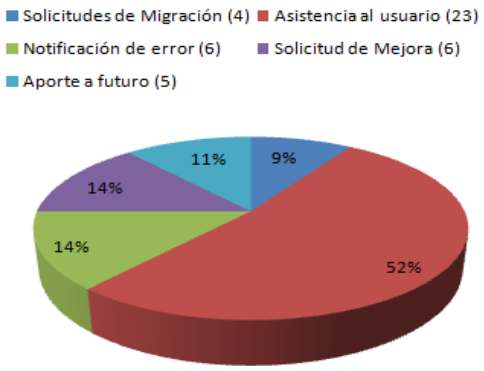


Figura 5: Distribución de Solicitudes al 12 de Marzo de 2012

6 Conclusiones

La Gestión del Conocimiento es una disciplina, no una tecnología que se pueda comprar y vender. Es una interacción que se produce entre procesos, personas y organizaciones, con la ayuda de la tecnología que le brinda soporte. Que esto resulte exitoso en una organización requiere de una cultura organizacional moderna, impulsada desde la dirección, que favorezca un ambiente estimulante para la colaboración, y a su vez brinde los métodos y herramientas para que sus miembros puedan compartir de manera eficiente su conocimiento explícito.

Para este trabajo, el enfoque está dado en el estudio de la Gestión del Conocimiento en la administración pública. Al respecto, se concluye que si bien las organizaciones públicas no deben preocuparse por su rentabilidad, sí deben hacerlo

por ajustarse a presupuestos y recaudaciones, y por sobre todo en superar problemas relacionados con factores del entorno y políticos.

Como caso práctico, se expuso el desarrollo de la Herramienta "Summun", sobre la cual se concluye que es un proyecto ambicioso pero a la vez realista y práctico, con metas que se fueron cumpliendo y resultaron en un sistema informático de carácter transversal al HTC. Se introdujo un concepto innovador: Gestión del Conocimiento; destacándose y difundiéndose sus ventajas a través de la capacitación; y llevado a modo práctico a través de Summun.

Para finalizar, puede afirmarse que la Gestión del Conocimiento debe implementarse como política de estado, entendiendo la importancia de la innovación tecnológica y conceptual, como la base indispensable sobre la que deben sustentarse las políticas públicas, fomentando la eficacia, equidad y transparencia en la gestión.

7 Trabajo futuro

La experiencia de los usuarios derivó en la necesidad de poseer un buscador integral que se abstraiga de la UI. Es decir, el buscador debe localizar palabras claves indistintamente del tipo de unidad, y mostrar todos los datos en una misma pantalla.

Se contempla el análisis y desarrollo de la etapa 3, planificando la generación de "digestos" que integren diferentes unidades para fines específicos. También vincular Summun con aspectos relacionados directamente con la gestión de calidad del Honorable Tribunal de Cuentas.

Finalmente se pretende lograr recomendaciones para el usuario, sustentadas en algoritmos del tipo "Acercamiento al vecino más cercano", actualmente aplicado en sitios como Amazon, Netflix, Reddit, entre otros [11].

Referencias

1. Catenaria, http://www.catenaria.cl/img/pdf/conocimiento.pdf
2. BusteloRuesta, Amarilla, Iglesias:Gestión del Conocimiento y Gestión de la Información. INFORAREA S.L., año VIII, n. 34. 2001
3. Seminario USAC, http://seminario1usac.wordpress.com/2011/05/08/business-intelligence/
4. Moore, Bresó Bolinches: El desarrollo de un sistema de gestión del conocimiento para los institutos tecnológicos.Revista Espacios, Vol.22. 2001
5. Lopez,Cabrales,Schmal: Gestión del Conocimiento: Una Revisión Teórica y su Asociación con la Universidad.Universidad de Talca, Chile.2005
6. Del Moral Bueno, Anselmo y otros: Gestión del Conocimiento. Paraninfo.2007
7. Wikipedia, Sociedad de la Información y del Conocimiento, http://es.wikipedia.org/wiki/Sociedad_de_la_informaci%C3%B3n_y_del_conocimiento
8. Rabinovitch, Jonas: Gestión del Conocimiento y Gobierno Electrónico: Mitos y Realidades. Departamento de Asuntos Económicos y Sociales (ONU), UNDESA, Noviembre 2009
9. Gob Prov. De Bs. As., http://www.gob.gba.gov.ar/legislacion/legislacion/l-10869.html
10. Flores,Coronel,Pardo, Groizard: "Summun. Conocimiento Organizacional". Presentación en Honorable Tribunal de Cuentas de La Provincia de Buenos Aires.2012
11. Wikipedia, Sistema Recomendador, http://es.wikipedia.org/wiki/Sistema_recomendador
12. Databases and Doctrine, http://symfony.com/doc/current/book/doctrine.html

Aplicación de los condicionales DHD en la herramienta del foro de Sakai

Alejandro R. Sartorio [1,2]
Marcelo A. Vaquero [2]
Guillermo Rodriguez [1,2]
Daniel Tedini [2]

[1] Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas, CIFASIS (CONICET-UNR-UPCAM), Bv. 27 de febrero 210 bis, 2000 Rosario, Argentina

[2] Centro de Altos Estudios en Tecnología Informática, CAETI, Universidad Abierta Interamericana, Sede Rosario, Ov. Lagos 944, 2000 Rosario, Argentina

{Sartorio, Guille}@fceia.unr.edu.ar, {Marcelo.Vaquero, Daniel.Tedini}@uai.edu.ar

Resumen. En este trabajo se brinda la información necesaria de la aplicación de un modelo de condicionales de la infraestructura de los contratos sensibles al contexto para el Dispositivo Hipermedial Dinámico (DHD). Partiendo de la implementación tecnológica en el entorno colaborativo SAKAI se muestra una forma de adaptar los servicios de la herramienta foro a la información de contexto de los usuarios. De esta manera, se describen un diseño general de integración para ser utilizado en implementaciones de cualquier tipo de condicionales para los DHD.

Palabras Clave: Coordinación de Contratos – Sistemas sensibles al contexto – Dispositivo Hipermedial Dinámico – TIC.

1 Introducción

El actual contexto físico-virtual que se construye a partir de la utilización de las Tecnologías de la Información y Comunicación (TIC) posibilita a los sujetos ser partícipes de redes sociotécnicas conformadas por una multiplicidad de componentes y relaciones, que se configuran y reconfiguran por las diversas interacciones en función de una gran diversidad de requerimientos. En este sentido, el Programa interdisciplinario de I+D "Dispositivos Hipermediales Dinámicos" [1], radicado en CIFASIS (CONICET-UNR-UPCAM), estudia la complejidad evidente de las mencionadas redes, integrando aportes de diversas disciplinas como informática, educación, ingeniería y psicología, entre otras.

Se conceptualiza como Dispositivo Hipermedial Dinámico (DHD) [2] a la red heterogénea conformada por la conjunción de tecnologías y aspectos sociales que

posibilitan a los sujetos realizar acciones en interacción responsable con el otro para investigar, aprender, dialogar, confrontar, componer, evaluar, bajo la modalidad de taller físico-virtual, utilizando la potencialidad comunicacional, transformadora y abierta de lo hipermedial, regulados según el caso, por una "coordinación de contratos" [3].

Funcionalmente, el DHD es conceptualizado como sistema complejo [4], en el cual los participantes realizan acciones de participación mediadas por diversas tecnologías. Estas interacciones se efectúan por medio de servicios provistos por herramientas específicas agrupadas según el espacio colaborativo utilizado. Además, se busca que los efectos de dichas acciones estén condicionados por la información de contexto de los participantes que la involucran.

Para esto, tecnológicamente el DHD está provisto por el agregado de una pieza de software diseñada para la inyección de propiedades de coordinación de contratos sensibles al contexto [5]. Esta propiedad se logra a través de la implementación de contratos [6] con mecanismos de coordinación y componentes de sistemas sensibles al contexto [7].

La utilización de reglas es parte esencial en la implementación de las acciones que contienen los contratos y las tareas de coordinación. A su vez, las estructuras de las reglas contienen condicionales donde se establece parte de la lógica de adaptación a los requerimientos funcionales del los DHD. En este trabajo se retoman los tres tipos de condicionales lógicos diseñados con el propósito de representar valores de verdad que dependan de la información de contexto de usuarios [19]. De esta manera se pretende diseñar una estructura conceptual que permita implementar estos condicionales en el marco tecnológico del DHD.

Tras esta introducción, en la sección 2 se identifican los elementos tecnológicos del DHD teniendo en cuenta su relación con los condicionales. Luego, en la sección 3, se presentan los tres tipos de condicionales con sus principales características y modelos de integración dentro del *framework* SAKAI (http://www.sakaiproject.org) [8]. En la sección 4 se describe un modelo conceptual genérico de condicionales junto a un ejemplo de integración. Por último, en la sección 5 se presentan las principales conclusiones generales.

2. El uso de contratos en el DHD

El uso de contratos en el DHD parte de la noción de Programación por Contrato (*Programming by Contract*) de Meyer [6] basada en la metáfora de que un elemento de un sistema de *software* colabora con otro, manteniendo obligaciones y beneficios mutuos. En nuestro dominio de aplicación consideraremos que un objeto cliente y un objeto servidor "acuerdan" a través de un contrato, representado con un nuevo objeto, que el objeto servidor satisfaga el pedido del cliente, y al mismo tiempo el cliente cumpla con las condiciones impuestas por el proveedor. De esta manera, las decisiones de comportamiento de los servicios se verán influenciadas por el valor de verdad de las instancias de los condicionales que integren al contrato.

Como ejemplo de la aplicación de la idea de Meyer [9] en nuestro dominio de tecnologías de la información y comunicación planteamos el escenario en que: un usuario (cliente) utiliza un servicio de edición de mensajes (servidor) a través de un contrato que garantizará las siguientes condiciones: el usuario debe poder editar aquellos mensajes que tiene autorización según su perfil (obligación del proveedor y beneficio del cliente); el proveedor debe tener acceso a la información del perfil del usuario (obligación del cliente y beneficio del proveedor).

A partir de la conceptualización de contratos se propone una extensión por medio del agregado de nuevas componentes para instrumentar mecanismos que permitan ejecutar acciones dependiendo del contexto. En aplicaciones sensibles al contexto, el contexto (o información de contexto) es definido como la información que puede ser usada para caracterizar la situación de una entidad más allá de los atributos que la definen. En nuestro caso, una entidad es un usuario (participantes, coordinadores, etc.), lugar (casa, universidad, parque, etc.), recurso (impresora, fax, etc.), u objeto (archivos de texto, fotos, videos digitales, etc.) que se comunica con otra entidad a través del contrato.

En [9] se propone una especificación del concepto de contexto partiendo de las consideraciones de Dourish [10] y adaptadas al dominio de las tecnologías de la información y comunicación, que será el punto de partida del actual trabajo. Contexto es todo tipo de información que pueda ser censada y procesada, a través de una aplicación, que caracteriza a un usuario o entorno, por ejemplo: intervenciones en foros, participaciones en wikis, habilidades, niveles de conocimientos, direcciones *ip* conectadas, cantidad de usuarios conectados, fechas y horarios, etc.

En términos generales, la coordinación de contratos es una conexión establecida entre un grupo de objetos influidas por condicionales que representan parte de la lógica de adaptación. Cuando un objeto cliente efectúa una llamada a un objeto servidor, el contrato “intercepta” la llamada y establece una nueva relación teniendo en cuenta el contexto del objeto cliente, el del objeto servidor, e información relevante adquirida y representada como contexto del entorno. Los condicionales de las reglas representarán diferente tipo de información de contexto con distinto grado de representación y abstracción, donde se requieren mecanismos de inferencias basados en la recolección, representación y simulación.

A través de un diagrama UML se definen las clases utilizadas en la implementación de los condicionales dentro de las reglas de los contratos. La Figura 1 describe los elementos y relaciones relevantes en la compocisión de condicionales.

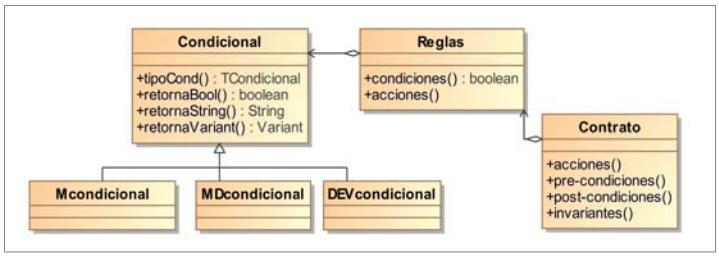


Fig.1. Elementos y relaciones relevantes en la creación de condicionales.

Dentro del mecanismo de configuración de condicionales, el contrato presenta en su interfaz métodos para la ordenación de acciones y reglas; además de las que cubren las propiedades inherentes a la definición original de contrato sobre la configuración de *pre-condiciones*, *post-condiciones* e *invariantes*. De esta manera se representa a las reglas anteriormente mencionadas como una clase de agregación del contrato. Así se determina una nueva clase para la representación de los condicionales. Las reglas contienen referencias a las acciones de los contratos por medio de la interfaz *acciones.*

Se decide representar a los condicionales como objetos de primera clase con el propósito de establecer un nuevo grado de abstracción que permitirá conectar a los contratos a subsistemas externos que le proporcionen nuevos mecanismos de adaptabilidad, dinamismo e interpretación.

De esta manera, teniendo en cuenta las experiencias de diseño e implementación del uso de condicionales [11] [12] [13] se extienden al objeto condicional en tres tipos diferentes. Cada uno de ellos hereda la interfaz *Condicional*, encargada de establecer las presentaciones para el tipo de dato necesario en las reglas.

El primer paso, es lograr la construcción de las reglas del contrato y que los condicionales representen criterios de decisiones sobre aspectos relevantes de los procesos didácticos, investigativos, de producción y/o de gestión mediatizados por un DHD; por ejemplo: un participante puede adquirir un servicio determinado de una herramienta a partir de la evaluación de una condición representada como condicional de una regla.

En general, a estas reglas debemos diseñarlas con el cuidado de no incorporar redundancias, ambigüedades o incoherencias; tanto entre las propias reglas de un contrato como con otras reglas implícitas que se desprenden de los servicios. Entonces, definimos a las *Reglas* del contrato como un conjunto de condiciones, acciones y prioridades. La condición es un expresión booleana sobre relaciones (mayor, menor, igual, distinto, etc.) entre parámetros y valores concretos. Las acciones conforman un conjunto de asignaciones de valor a otros parámetros también definidos por el tipo de regla. Algunos de los parámetros de las acciones deben ser "métodos de cálculo" que permiten cambios en el comportamiento de los servicios en los cuales estas reglas son aplicadas. La prioridad permite simplificar la cantidad de reglas que se deben escribir: en lugar de la escritura de una regla para cada

combinación de posibilidades de los valores de los parámetros, se asegura que dos reglas no puedan ser ejecutadas simultáneamente. Por ejemplo: el usuario podría escribir una prioridad baja para todas las reglas y luego con prioridades altas ir identificando las excepciones para el caso configurado inicialmente. En síntesis, las reglas son ejecutadas mediante un orden de prioridades.

Entonces, las reglas forman parte de un mecanismo de agregación encargado de la composición de diferentes tipos de condicionales: *Mcondicional*, *MDcondicional* y *DEVScondicional*, que se comportan de manera similar teniendo en cuenta diversos modelos de integración que explicitaremos en las secciones siguientes.

A continuación, se analizarán las estructuras de tres tipos de condicionales con el propósito de abstraer características de su conformación. En particular se identificarán tipos de elementos, patrones, sub-estructuras y relaciones que sean necesarias para la composición con los contratos.

3. Los tipos de condicionales

Retomando la sección anterior y de manera general a través de los elementos utilizados en el diseño de la Figura 1 para la creación de condicionales, se identifica como la primer característica de diseño, a la estructura que define los diferentes tipos de condicionales para el DHD [16,17,18.19].

Luego, partiendo de las necesidades adaptativas de los contratos y ante determinados casos de uso, se diseñaron diferentes implementaciones de condicionales. Estas diferencias tienen que ver con cuestiones de diseño en los que algunos aspectos pueden ser generalizados, posibilitando una representación más genérica de los condicionales.

Partiendo de los avances para la implementación del modelo conceptual de [19] se propone una método de inyección del modelo en la herramienta foro del *framework* Sakai. De esta manera, en la Figura 2 se representa un diagrama de las principales clases que integran la herramienta original del foro. En este caso, se toma la clase *BaseDiscussionMessage,* que extiende las operaciones de *BdDiscussionService*, para reimplementar el método *addDiscussionMessage*() encargado de ingresar el texto de una entrada de usuario al foro. A su vez, *addDiscussionMessage()* usa dos operaciones de *DiscussionMessageEdit* que utilizan servicios bases colaborativos extendiendo la infraestructura provista por las clases *Message* y *Edit*.

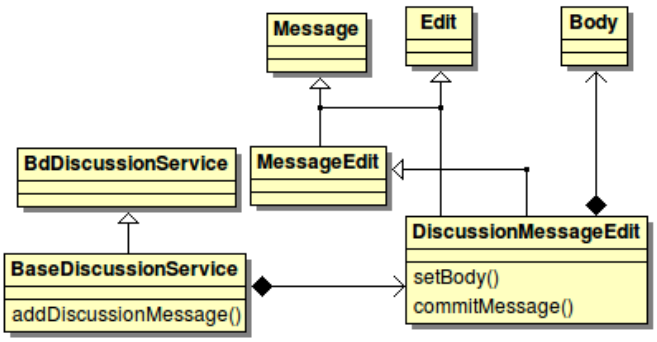


Fig. 2. Representación del los elementos de diseño de condicionales.

4. Inyección de los condicionalesDHD

En esta sección se presenta un diseño conceptual e información necesaria para la inyección de condicionales en una herrramienta del DHD.

En este caso, se propone un diseño de integración para conectar un subsistema de configuración (*Calculo*) para la instanciación de los condicionales de las reglas de contratos.

La Figura 3 establece el diseño propuesto para la implementación de condicionales. Se define un módulo para efectuar los cálculos finales que determinan el valor de verdad del condicional (*Calculo*). Otro módulo es el encargado de la recolección y toma de datos (*TomarDatos*), extendiéndose para los casos particulares donde es necesario contar con estructuras de datos (*Estructuras*) conteniendo métodos que implementan cada una de ellas.

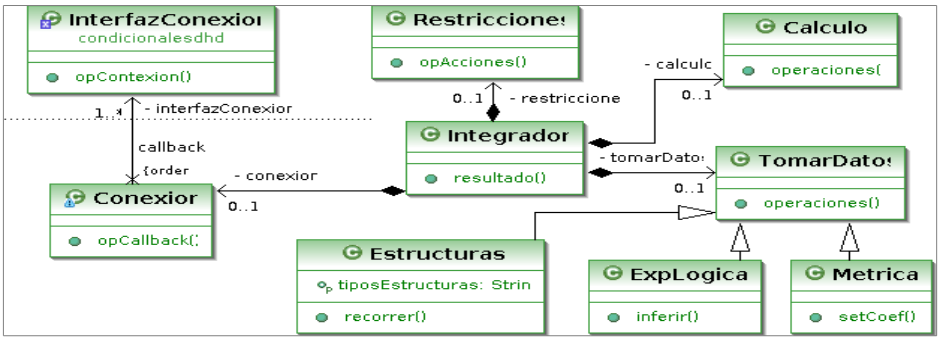


Fig. 3. Modelo de diseño conceptual de condicionales para contratos sensibles al contexto.

Además, un módulo aparte se configura para describir todas las restricciones que

debe cumplir el condicional (*Restricciones*), teniendo en cuenta su utilización dentro de las reglas de los contratos, con el propósito de no incurrir en contradicciones o inconsistencias en relación a las precondiciones, poscondiciones e invariantes.

Las conexiones con otros subsistemas, por ejemplo, el subsistema sensible al contexto representado, se encuentran encapsuladas en otro módulo de conexión (*Conexion*). De esta manera, se implementa un *"callback"* del método perteneciente a la interfaz de un subsistema externo.

4.1 Ejemplo

En la Figura 4 se muestra un ejemplo reducido de laboratorio para sistemas web colaborativo sensibles al contexto. Para esto exponemos un diagrama de secuencia que interpreta la ejecución de una regla del contrato que utiliza uno de los tipos de condicionales respetando el diseño original a través del objeto *BdDiscussionService*, y el modelo conceptual de integración a través de los objetos *Condicional*, *Metodo* y *TipoMetrica* [19]. De esta manera queda evidenciada la ventaja de contar con un modelo de diseño que facilite la manipulación de las reglas de los contratos sensibles al contexto.

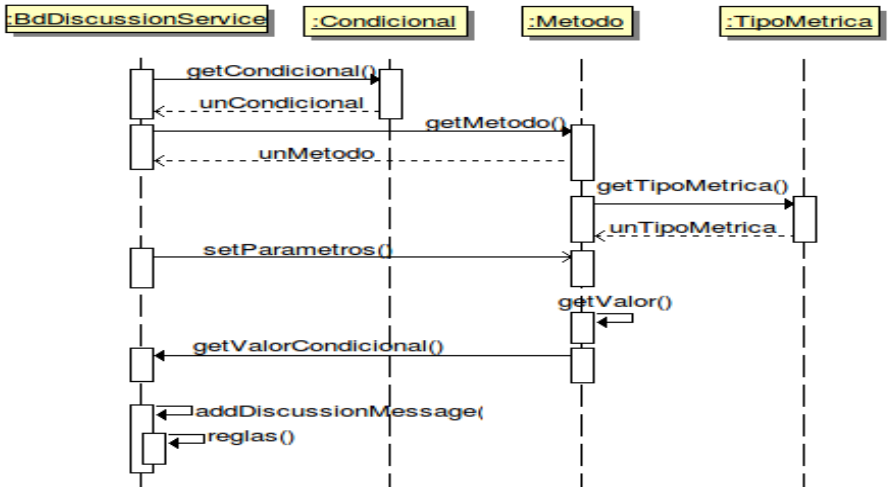


Fig. 4. Secuencia de invocación de los valores de verdad.

Luego, usaremos como referencias un fragmento de código descripto en el caso de estudio de la sección 5 de [5] para ejemplificar el impacto de esta propuesta. En este caso, se muestra que es necesario influir solamente sobre los conectores de la infraestructura de coordinación propuesta. La forma correcta de hacerlo es caracterizar la siguiente porción de código como una interfaz de conexión:

```
public CrdPartnerRules messageEdit_rules(string texto,Student c) throws
DiscussionException, CrdExFailure {return new CrdPartnerRules (this);}
```

Entonces, se debe agregar a esta función las siguientes lineas de códigos que posibilitarán cambiar las referencias de los condicionales de las reglas de los contratos:

```
Integrador unIntegrador = (Integrador) nuevoItegrador(); ...
unValorCondicional = unIntegrador.resultado(unDato, unCalculo);
Conexion unConexion=(Conexion) generarConexion (unValorCondicional);...
InterfazConexion unInterfazConexion = (InterfazConexion) nueva;...
unConexion.setConexion(unaInterfazConexion.metodoConexion());..
```

De esta manera se debe agregar estas líneas dentro del código fuente de la operación *addDiscussionMessage()* del la clase *DiscussionMessage.* Para este ejemplo particular, se reemplaza la línea 5. Este procedimientos se denomina inyección de condicionalesDHD y se puede aplicar en reemplazo de cualquier porción de código que representa la invocación de un servicio base [13,20] de las herramientas del *framework* Sakai.

```
public DiscussionMessage addDiscussionMessage(String category, String
subject, boolean draft, String replyTo,List attachments, String body)
throws PermissionException
{
1.      DiscussionMessageEdit edit=(DiscussionMessageEdit) addMessage();
                        ...
2.      edit.setBody(body);
3.      header.replaceAttachments(attachments);
                        ...
4.                      commitMessage(edit);
5.                      return edit;
6.              } // addDiscussionMessage
```

5. Conclusiones

Partiendo de los avances de diseño e implementación de una infraestructura para la integración de condicionales para los contratos sensibles al contexto del DHD, se brindaron información necesaria para la inyección en el servicio base de edición que implementa la herramienta foro del *framework* Sakai.

Como principal aporte de este trabajo se incorpora una extensión de las propiedades de los contratos sensibles al contexto para su adaptación en la capa de servicio de un framework web colaborativo particular.

Referencias

1. Programa I+D "Dispositivos Hipermediales Dinámicos", radicado en el Centro Internacional Franco Argentino de Ciencias de la Información y Sistemas (CIFASIS: CONICET-UNR-UPCAM). http://www.mesadearena.edu.ar. Directora: Dra. Patricia San Martín.
2. San Martin, P.: *Hacia un dispositivo hipermedial dinámico. Educación e investigación para el campo audiovisual interactivo*. Bernal: Universidad Nacional de Quilmes Editorial. 2008.
3. Sartorio, A. y Cristiá, M.: *Primera aproximación al diseño e implementación de los DHD*. XXXIV Congreso Latinoamericano de Informática, CLEI 2008. 2008.
4. Gell-Mann, M.: *El quark y el jaguar. Aventuras en lo simple y lo complejo*. Barcelona: Tusquets. 1995.
5. Sartorio, A. and Cristiá, M.: *First Approximation to DHD Design and Implementation*. Clei electronic journal, Vol.12, Nº 1. 2009.
6. Meyer, B.: *Applying Design by Contract.* IEEE Computer Society Press, Volume 25 Issue 10. pp. 40-51. 1992.
7. Dey, A.K., Salber, D. and Abowd, G.: *A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications, anchor article of a special issue on Context-Aware Computing*. Human-Computer Interaction (HCI) Journal, Vol. 16 (2-4). pp. 97-166. 2001.
8. Sartorio, A.: *Un modelo comprensivo para el diseño de procesos en una Aplicación E-Learning.* XIII Congreso Argentino de Ciencias de la Computación. CACIC 2007. 2007.
9. Sartorio, A. y San Martín, P.: *Sistemas Context-Aware en dispositivos hipermediales dinámicos para educación e investigación, en San Martin, P.: Hacia un dispositivo hipermedial dinámico. Educación e investigación para el campo audiovisual interactivo*. Bernal: Universidad Nacional de Quilmes Editorial. 2008.
10. Dourish, P.: *What we talk about when we talk about contex*t. Personal and Ubiquitous Computing, Vol. 8, Nº 1. pp. 19-30. 2004.
11. Sartorio, A., Rodríguez, G. y Vaquero, M.: *Condicionales DEVS en la coordinación de contratos sensibles al contexto para los DHD. XVI Congreso Argentino de Ciencias de la Computación, CACIC 2010. 2010.*
12. Sartorio, A.: *Los contratos context-aware en aplicaciones para educación e investigación, en San Martin, P.: Hacia un dispositivo hipermedial dinámico. Educación e investigación para el campo audiovisual interactivo*. Bernal: Universidad Nacional de Quilmes Editorial. 2008.
13. Sartorio, A., Rodríguez, G. y Vaquero, M.: *Investigación en el diseño y desarrollo para el enriquecimiento de un framework colaborativo web sensible al contexto.* XIII Workshop de Investigadores en Ciencias de la Computación, WICC 2011. 2011.
14. Rivera, M.B., Molina, H. y Olsina, L.: *Sistema Colaborativo de Revisión para el soporte de información de contexto en el marco C-INCAMI*. XIII Congreso Argentino de Ciencias de la Computación, CACIC 2007. 2007.
15. Olsina, L. and Rossi, G.: *Measuring Web Application Quality with WebQEM*. IEEE Multimedia, 9(4). pp. 20-29. 2002.
16. Rodríguez, G.: *Desarrollo e implementación de métricas para el análisis de las interacciones del Dispositivo Hipermedial Dinámico*. Jornadas Argentinas de Informátic, JAIIO 2010. 2010.
17. PowerDEVS 2.0 Integrated Tool for Edition and Simulation of Discrete Event Systems. Desarrollado por: Esteban Pagliero, Marcelo Lapadula, Federico Bergero. Dirigido por

Ernesto Kofman. Disponible en: http://www.fceia.unr.edu.ar/lsd/powerdevs/index.html
18. Rodríguez, G.: *SEPI-DHD: Herramienta integrada para el Seguimiento y Evaluación de los Procesos de Interactividad del DHD*, en: San Martín, P. y Traversa, O.: *El Dispositivo Hipermedial Dinámico Pantallas críticas: I+D+I para la Formación Superior en Crítica y Difusión de las Artes*. Ciudad Autónoma de Buenos Aires: Santiago Arcos. 2011.
19. Sartorio, A., Rodriuez, G., Vaquero, M.: *Hacia un diseño general de integración de condicionales para los contratos sensible al contexto del DHD*. CACIC 2012. 2012
20. Dagger, D., O'Connor, A., Lawless, S., Walsh, E., Wade, V.P.: *Service-Oriented E-Learning Platforms: From Monolithic Systems to Flexible Services*. Internet Computing, IEEE (Volume:11 , Issue: 3). Page(s): 28 – 35. 2007

Framework for Brain Computer Interface implemented to control devices in the context of home automation

Jorge Ierache[1,2], Gustavo Pereira[1] , Enrique Calot[2] , Juan Iribarren[1]

Instituto de Sistemas Inteligentes y Enseñanza Experimental de la Robótica (ISIER)[1]
Laboratorio de Sistemas de Información Avanzados Facultad de Ingeniería Universidad de Buenos Aires[2]

ISIER, Facultad de Informática Ciencias de la Comunicación y Técnicas Especiales
Universidad de Morón, Cabildo 134, (B1708JPD) Morón, Buenos Aires, Argentina
54 11 5627 200 int 189
jierache@yahoo.com.ar

Abstract. This paper presents the device controlling using brain machine interface (BMI), through the development of a Framework implemented for the acquisition, configuration and control in the context of home automation.

Key words. Domotic, Brain Machine Interface, Bio-Electrical Signal, Human Machine Interfaces, Neuroengineering

1. Introduction

In the field of emerging interfaces presents Brain Computer Interface (BMI, also known as BCI for its acronym in English), it facilitates communication between mental or cognitive functions created from the brain of a person, capturing the electrical signals to be processed, classified, and communicated with specific devices or applications. It is very interesting how applications employing BMI interfaces have increased during the last two decades [1], from controlling the lights on and off, using wheelchairs, computer control [2], movements in space [3], to video games [4]. In science the initial interest in using BMI presented since 1973 [1], among the first publication in the field of research in BMI were conducted in the nineties 1990 [5] and 1991 [6]. The application of controlling by bio-signals systems, robots, applications, games and other devices, presents a new approach to open the doors for interaction between humans and computers in a new dimension, which specifically exploit electrical biopotential registered user through: the electro-myogram, the EEG and electro-oculogram, which are electrical biosignal patterns generated by muscle activity, the welcome and the user's eyes.

Researching of BMI interfaces is developed in a multidisciplinary scientific field due to their medical, electrical and electronic signal processing components, neuroscience and applications like computing, home automation to robotics and Entertainment [7]. Several papers were presented: first, they used intracranial electrodes implanted in the motor cortex of primates [8], [9]. The noninvasive's human works used electroencephalography signals (EEG) applied to mental exercises commands, such as moving a computer cursor [10] , [11] based on the use of Brain- Machine Interface (BMI) . Millan et. al. [12] demonstrates how two people can move a robot using a simple electroencephalogram based on three recognizing mental states, which are associated with the robot command. The work Saulnier et. al. [13] focused on controlling the speed of a robot to extend its application to infer the user's stress level , and from this influence the social behavior of domestic robots , in this case a robot vacuum cleaner. The seminal work of Millan et. al. [12], used as an unique biosignal the EEG, based on two people working to give support in robot navigation , unlike the latter, our paper presents the preliminary result using a BMI of low cost, used in works like Saulnier et . al. [13] which includes the biosignals for the electroencephalogram, electro-oculogram and electromyogram. Unlike Saulnier's et. al. [13] work , which

implements a speed control based on electromyogram and infers the state of stress of the user through the electroencephalogram , our initial work focused on executing commands using a NIA BMI [14] to navigate of a robot [15] and is currently in the control of artifacts in the context of home automation. [16]. Control devices , moving robots or facilitate the implementation of devices for disabled without applying manual controls and gain control only through mental activity fascinated researchers, while achieving a plasticity with a BMI of time required by the user, on our experiences to facilitate employment to a user with minimal training was developed for auto focus control [15] in order that the Lego NXT robot [17] is guided by the use of a BMI-NIA, to accomplish a navigation pattern, improving mental control times, slightly surpassing the manual control, in performance tests of the same navigation pattern [18] , [19] . In 2011 researchs we experience the remote controlling of a Lego NXT robot [20] via the Internet with the implementation of biosignals with NIA BMI [14]. In subsequent works [21], [22] navigation control using Emotiv's BMI navigation was developed. [23], which is detailed in the next section.

2. Non-invasive brain machine interface

On biosignals, EEG is the registry of the electrical activity generated by the neurons within the brain which is obtained through the skull by the use of electrodes placed on the surface of the head. Neuronal electrical activity is composed of slow waves which originate in synaptic activity of cortical neurons. Obtaining bioelectric signals is performed on the scalp using surface electrodes, which give the name of electroencephalogram (EEG). As to the surface electrode types are distinguished the ones attached, that consist of small metal discs which are fixed with a conductive paste giving very low contact resistance. Contact also exists, consisting of small tubes of chloridized silver threaded plastic holder containing at its end a pad wetted with saline is attached to the skull with elastic bands which are connected via alligator clips. Finally we have a mesh helmet composed by electrodes attached to a elastic helmet which are more comfortable for use and have a high positioning accuracy. The Emotiv Epoc's electrodes [23] are subject to a malleable plastic arms that ensure the proper location and have a pad soaked in a salt solution into each contact to allow conduction.

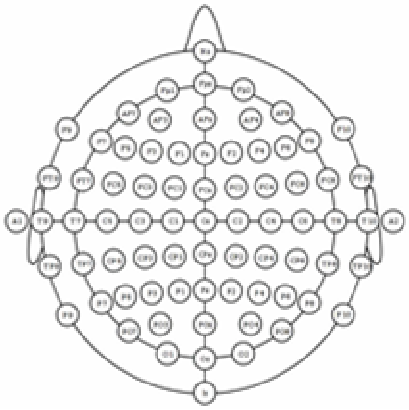


Figure Nº 1 "Electrode system 10-20"

The international 10-20 electrode arrangement (Figure N º 1) is the most used and developed one in order to ensure standardization for an individual studies. The system consists of letters and numbers to identify contact points. The letters identify the lobe and the number, the location within the hemisphere. The letters F, T, C, P and O stand for frontal, temporal, central, parietal and occipital respectively, (the letter C is used to identify the central horizontal line does not refer to any lobe). Even numbers correspond to the electrodes of the right hemisphere and left odd. Z subscripts are used to identify the vertical center line of electrodes. The arrangement of the electrodes on the Emotiv Epoc helmet, fit the 10-20 system, but only fourteen contact positions are used (Figure N ° 2a) plus a pair of reference (Common Mode Sense-CMS-and Driven Right Leg-DRL-) on each side, behind the ear or above it. In addition to the electrodes, the Emotiv Epoc helmet (Figure N º 2b) contains a gyroscope consisting of two accelerometers that provide information on the movements that user performs with his head and a wireless transmitter by which links with the USB receiver connected to the computer, all powered by a rechargeable battery via USB.

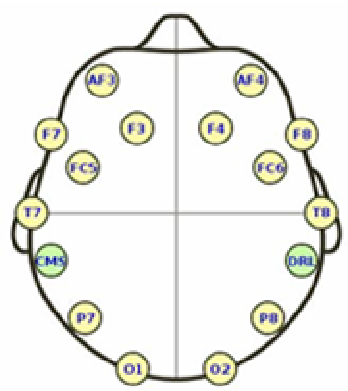


Figure Nº 2a. "EMOTIV EPOC helmet electrodes"

Figure Nº 2b. "EMOTIV EPOC helmet"

From the API and the software included in the development kit, the level of contact of the electrodes, the motion of the gyroscope, the wireless signal strength and battery charge can be monitored. Emotiv Development Kit has a set of libraries that allow communication with the helmet, the API for developers and the Emotiv engine. The Emotiv engine is the base component for the detection and interpretation of the signals from the electrodes that are located in the helmet and information captured by the EEG. It is also responsible for monitoring the battery state, the intensity of the wireless signal, the record of the connection time and to train recognition algorithms for expressive and cognitive modes, subsequently applying optimizations to each of them.

The Emotiv EPOC BMI (Brain Machine Interface), articulated with its SDK (Figure N º 3) consists of a control panel to create the user and log the profile, it also helps to visualize the connection status of the sensors and represents different record patterns (expressive, affective and cognitive). The expressive pattern can be viewed through an avatar in which signs of facial expression (blink, wink left, wink right, look left, look right, brows move up, move eyebrows down, smile, grit your teeth) can be trained. The affective pattern verifies different moods that are happening in a certain time (concentration, instant arousal, excitement among others). The cognitive pattern allows the training of an action on the basis of a thought, on which you can train up to thirteen actions, six of which are directional movements (push, pull forward, left, right, up and down), six rotational (rotation in the direction of clockwise, counter-rotating in the direction of clockwise, rotate left, rotate right, forward, backward) and an imaginary one that is disappearing. Other tool on the SDK is the Emokey (figure Nº 4), which allows Emotiv to link an action with any key and thus to function as an interface to any application.

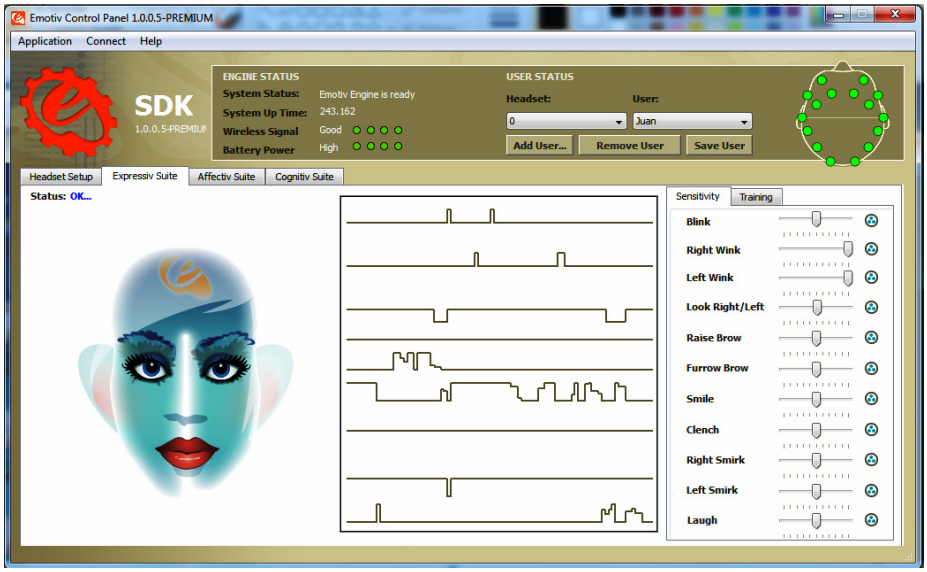


Figure Nº 3 SDK-Emotiv

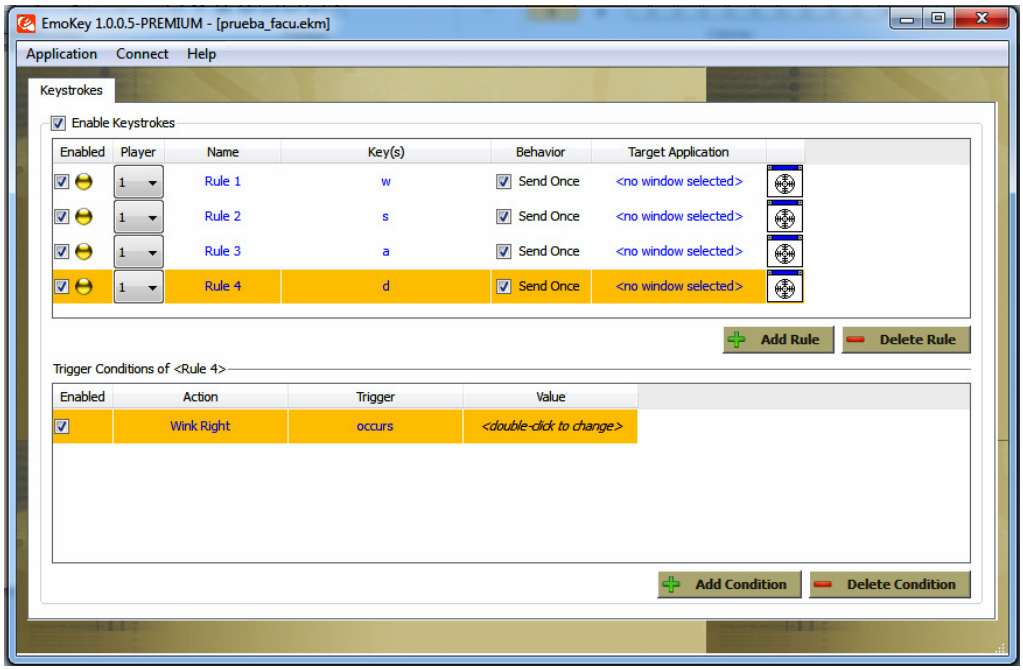


Figure Nº 4 "Emokey"

3. Framework to control artifacts in the context of home automation

The framework to control artifacts was implemented to allow control of devices through BMI-Emotiv [23]. Functional restrictions were raised as to implement the control with the least amount of commands to facilitate learning and control by users, in this order only two commands are used to control devices, one is dedicated to action selection (change the tv channel, change the volume, turn on or off, change the temperature, mode of air conditioning, etc.) and another is dedicated to their execution. For the integration of communications to replace the IR remote control USB-UIRT IR transceiver is used [24], which was also applied to capture the command to be incorporated into the Framework that will execute the using the BMI. In sum, through the transceiver commands will be captured from any IR home device remote control (TV, DVD player, air conditioning, etc.) then the UI allows distribution of the buttons in a more comfortable way for the user and using a friendly interface and iconography to simplify identification. Figure No. 5 shows conceptually the control application in the context of the automation.

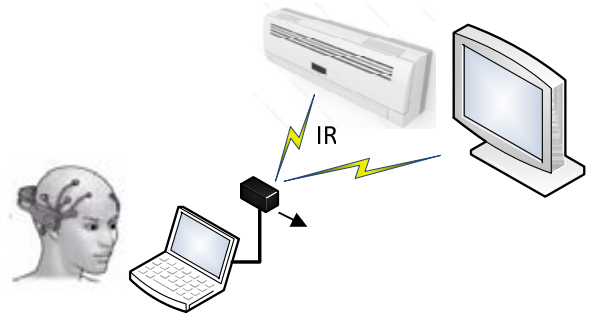


Figure N º 5 "Application automation with Emotiv/USB-UIRT"

Developed Framework will facilitate disabled persons to control devices, as well as the interaction and control of such users with remote devices on site. Over these facilities a simple profile is determined for managing a device, and in first place characterize and associates the control to selecting a command based on the detection of muscle signals, in this case through a slight movement of the eyelids. Second, executing high level commands to a device, in this case worked using alpha brainwaves. The framework was developed in C#. Class diagram is presented in Figure No. 6, it consists of two sections, the first one is configuring the layout of commands where you learn (catcher) and set the IR codes of the remote control device chosen (Class Configuration) and the second one where the user interacts with the devices (Class MainForm) by USB-UIRT API Managed Wrapper [25].

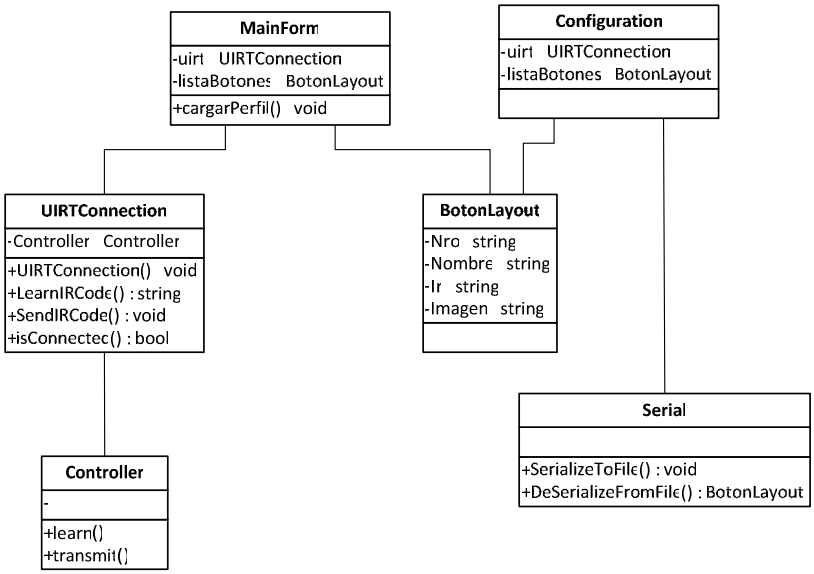


Figure N º 6 "Framework class diagram"

Figure Nº 7 shows the diagram of components that make the application of the Framework, interaction with USB-UIRT interface is observed, through the MainForm class

with which it develops control devices and their interaction with the Emotiv transparently controlling through biosignals. using Emokey. Finally, Serial Class converts into a plain text acquired through its Serialize method using the command to control a device that corresponds to a hexadecimal value. To execute the command applies the Deserialize method that allows recovering from the plaintext in hexadecimal command to the USB-UIRT IR transmitted via the device.

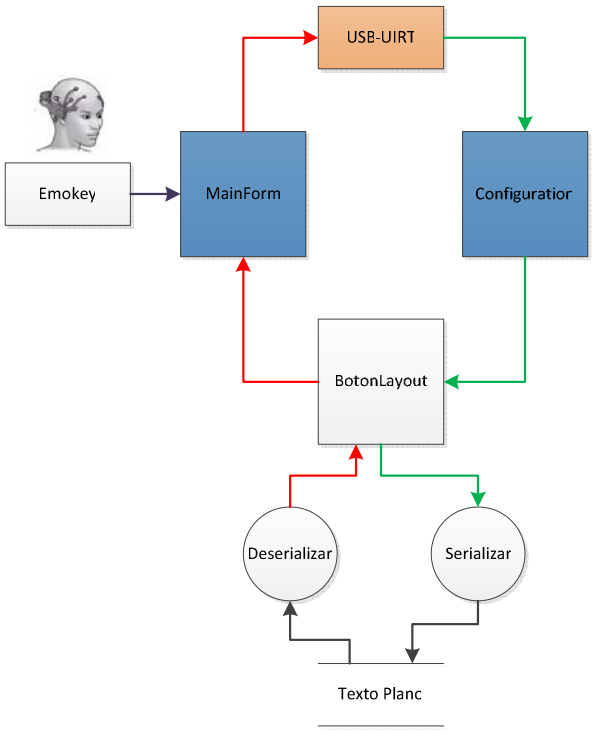


Figure N º 7 "Component diagram"

In the Configuration section the user selects the button that performs the action with the possibility of building an specific layout distribution, places the button name, choose an icon that matches and then copies the hex value of the command on the remote control using the Learn button and with USB-UIRT device connected. Once finished the layout configuration, it is stored in a disk file to be recovered in future runs of the application (Figure N º 8).

Figure Nº 8 "Mapping controls configuration"

The second section (MainForm) is the simplified layout display was configured in the above mentioned section and it interacts with the end user of the application. By two command executions assigned in the Emokey (move right and run) the user will select the controls and runs according to your needs (Figure Nº 9), the layout are presents only the configurated buttons, also allowing to visually distribute commands for user comfort.

Figure Nº 9 "Command Execution"

4. Conclusions and future lines of research

Framework development, on the initial basis of robot control and its extension to control devices, with the ability to facilitate the capture and configuration of commands on the

devices demonstrated during tests [22] a stable behavior in the devices integrated control over Emotiv BMI.

Future lines of research will be implemented over extended control features as artifacts both robots infrared remote control (air conditioning, TV, and other devices), targeting both local devices and those located in remote sites by an unique cross-platform framework that integrates libraries for a lot of known devices, command profiles import and export and an assistant for training and familiarization with the Emotiv user.

5. References

[1]Hamadicharef, "Brain Computer Interface (BCI) Literature- A bibliometric study", in 10th International Conference on Information Science, Signal Processing and their Applications, Kuala Lumpur, 2010, pp. 626-629.

[2] P. R. Kennedy, R. Bakay, M. M. Moore, K. Adams, and 1. Goldwaithe, "Direct control of a computer from the human central nervous system," IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 2, pp. 198-202, June 2000

[3] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," Proceedings 629 of the National Academy of Sciences (PNAS), vol. 101, no. 51, pp. 17 849-17 854, December 2004.

[4] http://edugamesresearch.com/blog/tag/bci/, 2013.

[5] J. R. Wolpaw, D. J. McFarland, and G. W. Neat, "Development of an Electroencephalogram-based Brain-Computer Interface," Annals of Neurology, vol. 28, no. 2, pp. 250-251, August 1990.

[6] J. R. Wolpaw, D. McFarland, G. Neat, and C. Forneris, "An EEG-based brain-computer interface for cursor control," Electroencephalography and clinical Neurophysiology, vol. 78, no. 3, pp. 252-259, March 1991.

[7] M. A. Lebedev and M. A. L. Nicolelis, "Brainmachine interfaces: Past, present and future," Trends in Neurosciences, vol. 29, no. 9, pp. 536-546, September 2006.

[8]. J. Wessberg, C. R. Stambaugh, J. D. Kralik, P. D. Beck, M. Laubach, J. K., "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," Nature, vol. 408, pp. 361–365, 2000.

[9].M. A. L. Nicolelis, "Brain-machine interfaces to restore motor function and probe neural circuits," Nature Rev.Neurosci., vol. 4, pp. 417–422, 2003.

[10] R. Wolpaw, D. J. McFarland, and T. M. Vaughan, "Brain-computer interface research at the Wadsworth center," IEEE Trans. Rehab. Eng., vol. 8, pp. 222–226, 2000.

[11] J. del R Millán, "Brain-computer interfaces," in Handbook of Brain Theory and Neural Networks, 2nd ed, M.A. Arbib, Ed. Cambridge, MA: MIT Press, 2002.

[12] José Millán, Frédéric Renkensb, Josep Mouriñoc, and Wulfram Gerstnerb. Non-Invasive Brain-Actuated Control of a Mobile Robot by Human EEG. IEEE Trans. on Biomedical Engineering, Vol 51, June 2004.

[13] Paul Saulnier, Ehud Sharlin, and Saul Greenberg. Using Bio-electrical Signals to Influence the Social Behaviours of Domesticated Robots. HRI'09, 2009, USA.ACM 978-1-60558-404-1/09/03.

[14] http://www.ocztechnology.com/products/ocz_peripherals/nia-neural_impulse_actuator vigente julio 2013

[15] Ierache Jorge, Pereira Gustavo, Iribarren Juan, Sattolo Iris, "Robot Control on the Basis of Bio-electrical Signals" : "International Conference on Robot Intelligence Technology and Applications" (RiTA 2012) Gwangju, Korea on December 16-18, 2012. Series Advances in Intelligent and Soft Computing of Springer.

[16] Ierache., J, Pereira.,G, Sattolo., J, Iribarren Aplicación de interfases lectoras de bioseñales en el contexto de la domótica XV Workshop de Investigadores en Ciencias de la Computación 2013 Facultad de Ciencia y Tecnología Universidad Auto noma de Entre Ríos (UADER), ISBN: 9789872817961.

[17] http://mindstorms.lego.com/eng/Overview/default.aspx nxt vigente Julio 2013.

[18] Ierache, J., Dittler M., Pereira G., García Martínez R.,(2009) "Robot Control on the basis of Bio-electrical signals" XV Congreso Argentino de Ciencias de la Computación 2009, Universidad Nacional de Jujuy, Facultad de Ingeniería, ISBN 978-897-24068-3-9 ,pag 30.

[19].Ierache, J., Dittler, M. García-Martínez, R., "Control de Robots con Basado en Bioseñales". XII Workshop de Investigadores en Ciencias de la Computación WICC 2010: Universidad Nacional de la Patagonia San Juan Bosco, Calafate, Santa Cruz, Argentina. 2010, ISBN 978-950-34-0652-6, pag 641

[20] Ierache., J, Pereira.,G, Sattolo.,I , Guerrero., A, D'Altto J, Iribarren,. J. Control vía Internet de un Robot ubicado en un sitio remoto aplicando una Interfase Cerebro-Máquina". XVII Congreso Argentino de Ciencias de la Computación 2011, Universidad Nacional de La Plata, Facultad de Informática, ISBN 978-950-34-0756-1, paginas 1373-1382.

[21] Ierache Jorge, Pereira Gustavo, Iribarren Juan del articulo "Demostración de los resultados en la integración de Interfases Lectoras de Bioseñales aplicadas al Control de un Robot" VII Congreso Educación en Tecnología y Tecnología en Educación 2012 Universidad Nacional del Noroeste de la Provincia de Buenos Aires. UNNOBA, 2012, demos educativas. ISBN 978-987-28186-3-0.

[22] www.facebook/isierum

[23] http://www.emotiv.com/ vigente julio 2013.

[24].USB-UIRT: <http://www.usbuirt.com/> vigente Julio 2013.

[25] Jordan Zaerr USB-UIRT Managed Wrapper <https://github.com/JordanZaerr/Usb-Uirt-managed-wrapper>

Una aplicación de la Wikimedia Semántica

Marcela Vegetti, Horacio Leone

INGAR (CONICET - UTN), Avellaneda 3657, Santa Fe, Argentina
{mvegetti, hleone}@santafe-conicet.gov.ar

Abstract. Ontolog es una comunidad abierta virtual que trabaja colaborativamente para impulsar avances en el campo de la ingeniería ontológica y las tecnologías semánticas. Se encuentra alojada en una plataforma wiki que no posee las características necesarias para realizar búsquedas dentro de este cuerpo de conocimientos. Asimismo, la cantidad de información almacenada y la organización de la misma dificulta el acceso por parte de personas no familiarizadas con la estructura del contenido. Este artículo presenta una propuesta que incorpora semántica a las páginas de la wiki Ontolog y, además, provee una capa que permite el fácil acceso a la información relevante, reutilizando la información almacenada.

Keywords: ICOM; wikimedia semántica; ontolog

1 Introduction

El foro ontolog[1] es una comunidad abierta virtual dedicada a impulsar avances en el campo de la ingeniería ontológica y las tecnologías semánticas. Una plataforma wiki archiva las contribuciones de los miembros de la comunidad sirviendo de repositorio de conocimiento para la misma.

Una de las actividades impulsada por el foro Ontolog es la cumbre de ontologías, una serie anual de eventos que involucra a representantes de la comunidad de ontologías así como de las comunidades relacionadas con el tema elegido para la cumbre de cada año. Al igual que el resto de los proyectos del foro Ontolog, toda la información sobre las diferentes actividades que son llevadas a cabo durante cada cumbre anual son mantenidas en la infraestructura colaborativa de Ontolog.

La cantidad de información, y la forma en que está organizada, dificulta el acceso rápido a las partes más importantes de la misma para aquellas personas no familiarizadas con la estructura del repositorio. Es por esto que el comité organizador de la edición 2012 decidió que a partir de ese año debería existir, además de la información en la plataforma wiki, un sitio web que resuma la información más importante. En 2012, este desarrollo ha implicado la reescritura de parte del contenido, duplicando la información con el esfuerzo e inconvenientes que eso implica.

[1] http://ontolog.cim3.net/

Para evitar este inconveniente y teniendo en cuenta el objetivo que tiene Ontolog de impulsar la utilización de las tecnologías semánticas en las aplicaciones, surge la necesidad de implementar una plataforma que facilite el acceso, el reuso y la incorporación de semántica al contenido del foro Ontolog.

Para atender estos requerimientos, este trabajo presenta una propuesta, que se está implementando en una plataforma wikimedia semántica, denominada Ontolog PSMW. La ontología ICOM (Integrated Collaboration Object Model for Interoperable Collaboration Services)[2] es utilizada para definir la semántica del cuerpo de conocimiento y una capa de presentación sirve de máscara a la información almacenada, proveyendo un fácil acceso a la información relevante. El resto de este artículo se organiza como sigue. La sección 2 presenta una breve descripción de la organización de la serie de cumbres de ontologías y de su contenido, así como de la tecnología de la plataforma en la que está contenida. La sección 3 describe las características de la propuesta e ilustra cómo se ha aplicado al cuerpo de conocimiento de la cumbre de ontologías 2013. Finalmente, en la sección 4 se presentan las conclusiones y trabajos futuros .

2 Cumbres de ontología. Descripción y plataforma de soporte

La serie de cumbres de ontologías se viene desarrollando como iniciativa del foro Ontolog y el NIST (US National Institute of Standards and Technology) desde el año 2006. Cada año se selecciona un tema que es discutido durante una serie de eventos que se extienden alrededor de tres meses, desde enero a abril. Estos eventos incluyen teleconferencias semanales, discusiones en línea sobre el tema elegido, así como otras actividades particulares a cada año. Cada cumbre finaliza con un simposio presencial en el que se resumen los resultados de las diferentes actividades. Además, cada año se presenta en un comunicado el mensaje de los participantes de la cumbre para la comunidad de ontologías.

El tema elegido cada año es abordado desde diferentes facetas, que son analizadas y estudiadas en distintos espacios de discusión, denominados "Tracks". Cada uno de éstos está a cargo de uno o dos coordinadores responsables. Cada track debe organizar una o más conferencias virtuales, las cuales comprenden entre 3 y 6 presentaciones cortas sobre temas dentro del alcance del mismo, seguida de una sesión de preguntas y respuestas entre participantes y panelistas invitados. Estas reuniones virtuales están soportadas por una audio conferencia y una sala de chat. Todo el proceso de la cumbre es coordinado por un comité organizador integrado por los responsables de las diferentes actividades y dos coordinadores generales.

Cada track produce como resultado una página wiki que sintetiza la discusión. Los coordinadores del track son los responsables de crear y mantener dicha página en base a: i) las discusiones por mail que se promueven en la lista [ontology-summit-list], ii) las contribuciones que los participantes hacen en una página destinada a tal fin (Track community input), iii) las presentaciones en las conferencias virtuales que

[2] https://wiki.oasis-open.org/icom

el Track organiza y iv) los resultados de actividades específicas propuestas por los coordinadores del track.

La cumbre de este año ha puesto el foco en "La evaluación de las ontologías a lo largo de su ciclo de vida" y ha sido organizada en cuatro tracks: i) Aspectos intrínsecos de la evaluación de ontologías, ii) Aspectos extrínsecos de la evaluación de ontologías, iii) Construcción de Ontologías para cumplir con criterios de evaluación, y iv) Entornos de software para la evaluación de ontologías.

Como se mencionara en la sección previa, el repositorio del foro Ontolog se encuentra alojado en una plataforma wiki. La misma permite, a través de los denominados "purple numbers" (números púrpura), referenciar a párrafos específicos dentro de una página utilizando un código alfanumérico único dentro de la página. Desde otras páginas es posible referenciar a estos nodos utilizando el identificador del nodo, denominado nid ("node identifier") o número púrpura, en el URL de la página. Esta plataforma wiki se constituye en un importante repositorio del conocimiento que se genera colaborativamente en las cumbres de ontología. Sin embargo, a pesar de su utilidad, la información allí contenida no es fácilmente utilizable por herramientas externas. No es sencillo recolectar información que se encuentra distribuida en diferentes páginas, como por ejemplo, cuáles son las conferencias virtuales en las cuales ha participado alguna persona. A pesar de que la información en una wiki se encuentra estructurada (cada conferencia tiene su página con links a miembros que participaron en ella), su significado no es claro para una computadora, ya que no se encuentra representado de manera procesable por máquina. Por lo cual, este contenido no puede ser consultado utilizando las nuevas tecnologías de la Web Semántica.

Asimismo, el cuerpo de conocimiento almacenado en la wiki Ontolog es muy voluminoso y posee una organización que impide que personas no familiarizadas con su estructura puedan acceder fácilmente al mismo. Si tomamos, por ejemplo la página de alguna conferencia virtual, la misma contiene bastante información sobre ella: el tema a tratar, quienes son los coordinadores, los panelistas, una pequeña reseña de cada uno de ellos y el enlace a sus trasparencias, también se muestra la transcripción del chat (una vez que fue realizada la conferencia) e incluso las instrucciones para poder conectarse para participar de la misma (información que no es importante una vez finalizada).

3 Incorporación de Semántica y Reorganización del Contenido

Como se mencionara en la sección previa, el contenido alojado en la wiki Ontolog está siendo migrado a una plataforma wikimedia que incorpora características semánticas. Sin embargo, la sola migración a esta nueva plataforma no agrega semántica al contenido, así como tampoco implica la reorganización del mismo. A fin de cubrir estas necesidades se presenta una propuesta que incorpora semántica al contenido a través de la ontología ICOM y una capa que provee acceso al repositorio de información mediante una vista reorganizada de la misma. La Fig. 1 muestra los componentes principales de esta propuesta, los cuales serán introducidos a continuación.

PSMW (Purple Semantic Media Wiki) es una plataforma que, mediante las extensiones SMW (Semantic Media Wiki) [1] y PMWX (Purple mediawiki) [2], permite agregar semántica y acceso de "grano fino" al contenido de páginas wiki. El contenido del foro Ontolog está siendo migrado a esta nueva plataforma, creando el foro Ontolog PSMW, que posibilitará la incorporación de semántica al contenido.

Fig. 1. Componentes de la propuesta

A través de la extensión SMW es posible hacer explícita la información acerca de las relaciones existentes entre las páginas wiki. En la mediawiki estas relaciones se establecen con los links entre las páginas. Cada enlace establece cierta relación entre dos páginas, sin explicitar qué tipo de relación se trata. SMW permite caracterizar estos links mediante la definición de propiedades, de manera que el destino de un link se convierte en el valor de una propiedad definida para la página en la que está el enlace. De esta manera, es posible formular consultas semánticas sobre el contenido de las páginas. Para conseguir esto se requiere la definición de 4 tipos de páginas diferentes: propiedades, plantillas, formularios y categorías, en sus correspondientes espacios de nombre: Property, Template, Form y Category.

Las propiedades, junto con los tipos, son las formas básicas de introducir datos semánticos en la wikimedia semántica. Una propiedad se utiliza para anotar una porción de información en una página. Como ejemplo, considere la Fig. 2 en la cual se muestra el código utilizado en la wiki original y en Ontolog PSMW para mostrar una porción de texto en una página wiki. En el segundo caso, se puede observar que, los enlaces de las páginas a los organizadores de Ontolog han sido anotados con la propiedad ***hasOrganizer***.

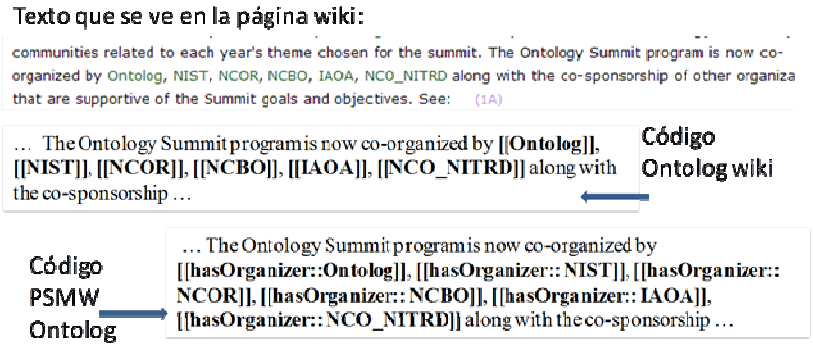


Fig. 2. Ejemplo de código para anotar texto en PSMW

Las plantillas son páginas wiki en el espacio de nombres Template cuyo contenido puede ser insertado en otra página. Una plantilla posibilita establecer la visualización

de los datos de una página y, además, permite convertir los datos en información semántica real ya que en ella es posible definir las propiedades de las páginas que la utilizan.

Los formularios permiten a los usuarios crear o editar páginas fácilmente. Se define un formulario por cada categoría y se asocia a una o más plantillas, las cuales permiten clasificar las páginas, que el formulario crea, bajo una Categoría. Las categorías son usadas como etiquetas para las páginas wiki e indican que una página pertenece a un tipo particular de página. Éstas son una característica de la wikimedia que permite el indexado automático de las páginas. Cada categoría establece un tipo de página diferente que se vincula con una plantilla y un formulario que permite la creación de sus correspondientes páginas.

Por su parte, la ontología ICOM, utilizada para definir la semántica de Ontolog PSMW, define un marco para la representación e integración de las actividades que se llevan a cabo en un entorno de colaboración [3]. Este vocabulario integra una amplia gama de actividades de colaboración, incluyendo y adaptando una serie de modelos que forman parte de normas y tecnologías existentes. ICOM posee una estructura modular para permitir la extensibilidad, a través del agregado de nuevos módulos. Los conceptos fundamentales y sus relaciones se incluyen en el núcleo (icom_core), mientras que los conceptos y las relaciones específicas para cada área de actividades de colaboración se definen en los módulos de extensión separados (icom_conf, por ejemplo). El subconjunto de los conceptos ICOM que son relevantes para la propuesta presentada en este artículo se muestran en la Fig. 3.

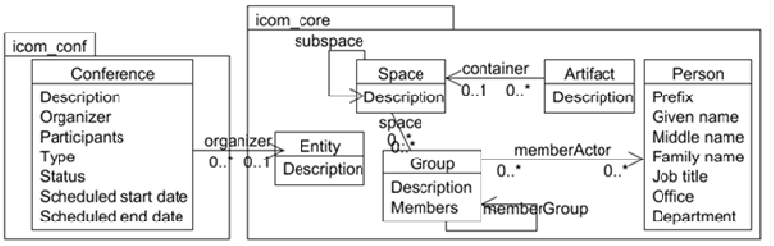


Fig. 3. Conceptos de ICOM utilizados en la propuesta

Una ***Entidad*** (Entity) es un objeto identificable que puede ser almacenado. Entity es la superclase de todas las otras clases definidas en ICOM (la relación superclase-subclase de Entity con los otros conceptos se ha omitido de la Fig. 3 para favorecer la claridad de la misma).

Un ***Grupo*** (Group) es una colección de individuos que comúnmente está asociado a un espacio de trabajo pero puede estar relacionado con más de uno (asociación ***space*** en la Fig. 3) y con cero o más subgrupos (relación ***memberGroup*** en la Fig. 3). Los individuos, Persona, (Person), miembros de un grupo, se vinculan a éste a través de la relación. ***memberActor***.

Un ***Espacio*** (Space) es un ámbito que define un marco para que los actores trabajen o colaboren. ***Espacio*** la representación concreta de un área de trabajo para una colaboración y puede estar vinculado con un grupo y con otros espacios subsidiarios.

El concepto ***Persona*** (Person) representa a un individuo que participan en una colaboración. Una persona puede ser miembro de un grupo así como tener su propio espacio de trabajo.

El resultado de las diferentes actividades colaborativas llevadas a cabo en los distintos espacios se representa como un ***Artefacto*** (Artifact) que se vincula a un espacio a través de la asociación ***container***. Un artefacto puede representar un documento o un conjunto de documentos relacionados que pueden estar incluidos en alguna página o almacenados en un repositorio.

Otro de los conceptos propuesto por ICOM es el de ***Conferencia*** (Conference), el cual representa una reunión que se lleva a cabo, de manera virtual o presencial, entre individuos. Una conferencia tiene un organizador, que puede ser un grupo o una persona, un conjunto de participantes, así como fechas de inicio y finalización tanto planificadas como reales.

La aplicación de estos conceptos al cuerpo de conocimiento relacionado con la cumbre de ontologías 2013 se muestra parcialmente en las Fig. 4 y 5. En particular, en la Fig. 4 se observa que la cumbre de ontologías del año 2013 se representa como un espacio de trabajo (***OntologySummit2013Space***) que está relacionado con el grupo (***OS2013Team***) que impulsa el desarrollo de las actividades de la cumbre. Este grupo, tiene a los coordinadores generales como miembros actores (relación ***memberActor***) y a su vez, está relacionado (relación ***memberGroup***) con los subgrupos que representan a los comités organizador (***OS2013OrgCommittee***) y asesor (***OS2013AdvCommittee***), a los patrocinadores (***OS2013Sponsors***) y a las instituciones organizadoras de la cumbre (***OS2013CoOrganizer***).

Por otra parte, la responsabilidad del espacio ***OntologySummit2013*** en la organización del simposio presencial, así como de las reuniones virtuales que no son específicas de un track, está representada también en la Fig. 4 mediante la relación ***organizer*** que se establece entre el espacio y las conferencias en cuestión. En la figura sólo se muestran dos de estas conferencias: ***Phase2PhaseSymposium*** y ***ConferenceCall_2013_04_25***. Asimismo, el comité organizador, tiene asociado un espacio de trabajo (***OS2013OrgCommittee***) y lleva adelante reuniones virtuales entres sus miembros, las cuales son representadas como conferencias (por ejemplo, ***OrganizingCommitte meeting n.07-Fri 2013.03.15***).

Como se mencionó anteriormente, en la cumbre de ontologías se llevan adelante diferentes tipos de actividades, las cuales pueden clasificarse en espacios de discusión denominados tracks y espacios transversales de trabajo, en los que se busca la realización de un trabajo concreto que, preferentemente, involucre aspectos discutidos en todos los tracks. Este año, por ejemplo, se concretaron el desarrollo de una encuesta, la redacción del comunicado, la organización de la biblioteca digital, el desarrollo del sitio web, además de la organización de conferencias tipo hackathon. En todos los casos, las actividades, son representadas como un espacio (icom_core_Space) subsidiario de ***OS2013Activities.*** En la Fig. 4 se muestran los espacios creados para representar las actividades de este año: ***SurveyDevelopment***, Hackathon&***Clinics***, ***CommuniqueDevelopment***, ***WebsiteDevelopment*** y ***CommunityLibraryDevelopment***. Como puede observarse en la Fig. 4 para Hackathon&***Clinics*** y ***CommuniqueDevelopment***, cada actividad: i) tiene un grupo

que la soporta, ii) puede proponer subactividades y iii) puede producir uno o más artefactos.

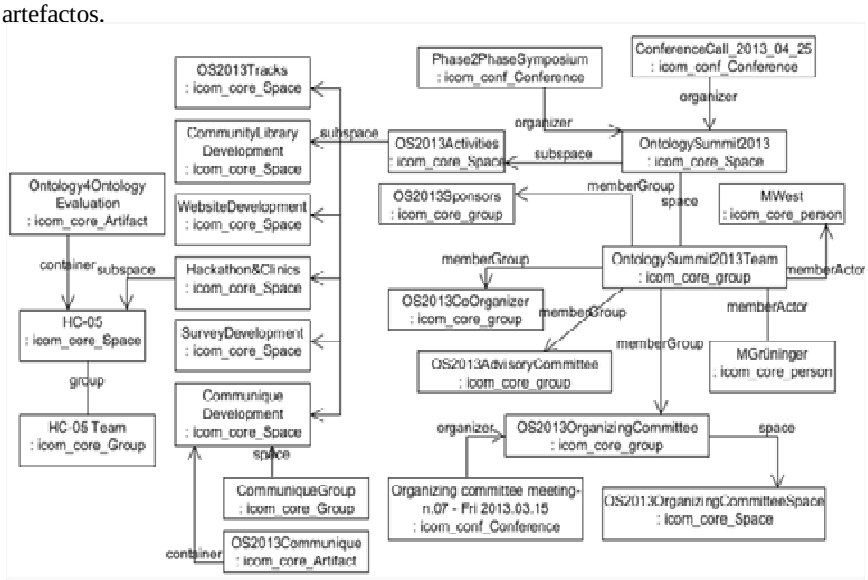


Fig. 4. Aplicación de los conceptos ICOM al contenido de la cumbre de ontologías de un año

Los espacios de discusión o tracks son un tipo particular de actividad que tiene como objetivo el estudio, análisis y discusión del estado del arte en un tema específico dentro del tema general de una cumbre particular. Cada uno de los tracks es representado como un espacio subsidiarios del espacio ***OS2013Tracks*** que se muestra en la Fig.4.

En la Fig. 5 cada Track, independientemente del tema en particular que aborde, está relacionado con: i) un grupo compuesto por sus coordinadores y sus panelistas, ii) un conjunto de conferencias virtuales, organizadas por el track y que abordan la temática bajo discusión desde la óptica de dicho track, iii) dos subespacios: uno en el que los participantes de la cumbre pueden hacer sus aportes a la discusión del track (OS2013TrackBCommunityInput) y otro, en el cual los coordinadores del track realizan la síntesis de la discusión llevada a cabo en el track (OS2013TrackBSynthesis).

Además, la Fig. 5 muestra el grupo asociado al espacio OS2013TrackB, con sus miembros que son los coordinadores (ToddS y TerryL) y los panelistas (HansP, MaryB, MeganK, JoaoPauloA, AmandaV, KeithS) que realizaron su presentación en alguna de las conferencias organizadas por el track (ConferenceCall_2013_01_24 , ConferenceCall_2013_02_28). Este espacio no tiene otros espacios subsidiarios.

En las dos conferencias ilustradas en la Fig. 5, así como en todas las conferencias semanales de cada cumbre, los panelistas realizan sus presentaciones y quedan como resultado de las mismas los siguientes productos o artefactos: i) las trasparencias de cada presentación, ii) la transcripción (en bruto y editada) de la sesión de chat y iii) la

grabación del audio de la conferencia. Todos estos artefactos son almacenados en un repositorio al que los participantes de la cumbre y cualquier interesado tienen acceso.

Los conceptos ICOM que se presentaron en la Fig. 3 se mapearon a PSMW para poder ser utilizados para anotar las páginas wiki de Ontolog. En este mapeo se crearon: i) ***propiedades*** para representar las relaciones y los atributos, ii) ***plantillas*** y ***categorías*** para representar los conceptos ICOM, y iii) ***formularios*** que permiten la creación de las páginas. En las diferentes plantillas se incorporaron consultas acerca de las propiedades definidas en las páginas.

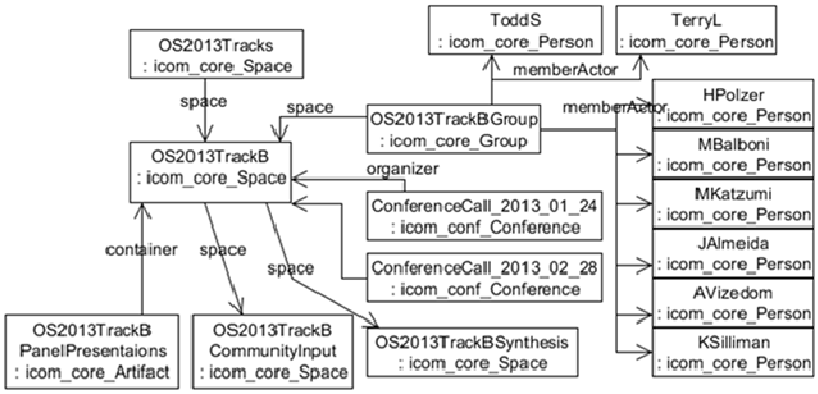


Fig. 5. Representación del Track B: Extrinsic Aspect of Ontology Evaluation

Para abordar el inconveniente de la complejidad del contenido y de la dificultad del acceso al mismo por parte de personas no familiarizadas con la estructura del repositorio, se propone incluir una capa de presentación sobre las páginas wiki que permita acceder a través de consultas y de sentencias de transclusión (incluir en una página wiki parte del contenido de otra página) al contenido que se pretende mostrar. En principio la tarea se focalizó solamente en el cuerpo de conocimiento asociado a la cumbre de ontologías de este año 2013. Sin embargo, la propuesta puede ser extendida sin mucho esfuerzo a cumbres de otros años y a otros eventos de Ontolog. Siguiendo la estrategia utilizada para el desarrollo del sitio web de la cumbre del año pasado, la propuesta divide el contenido de la cumbre en tres espacios: Actividades, Recursos y Entregables.

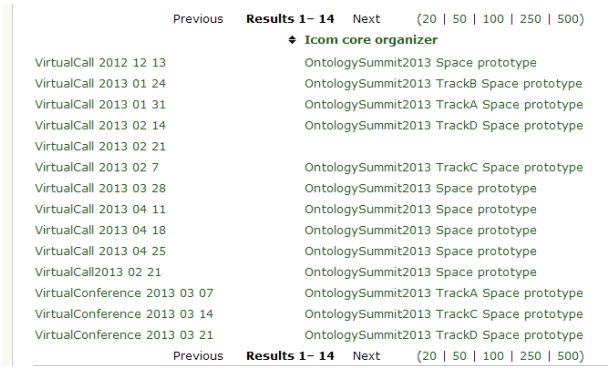
El espacio de Actividades presenta información sobre las diferentes actividades propuestas y posibilita el acceso a los respectivos espacios de trabajo. Los otros dos espacios, permiten a los usuarios acceder a la lista de las conferencias virtuales, la biblioteca digital, el archivo de la lista de discusión, en el espacio ***Recursos*** y a todos los artefactos que son resultado de las diferentes actividades que se llevan a cabo durante los tres meses de la cumbre en el espacio ***Entregables***.

Las páginas que forman parte de esta capa no han sido migradas desde la vieja plataforma sino que se crearon utilizando las propiedades, plantillas, categorías y formularios definidos al mapear los conceptos de ICOM a PSMW. Además, con el fin de reutilizar el contenido existente en las páginas migradas, se hace uso de las

consultas que provee la extensión SMW y de la capacidad de transclusión provista por la wikimedia. El objetivo de la transclusión es obtener los últimos datos referenciales. En otras palabras, si se cambia la página de referencia, a continuación, la información incluida se actualizará automáticamente. PSMW soporta actualmente transclusión de toda una página, o de una sección de la página (a través de los números púrpura).
Con esta propuesta se ha incorporado semántica al cuerpo de conocimiento relacionado con la cumbres de ontologías 2013, lo cual permite realizar consultas como la planteada al final de la sección 2 de ese artículo: ***¿Cuáles son las conferencias virtuales en las cuales ha participado una determinada persona?*** La misma puede realizarse con la siguiente función #ask, en la cual se consulta por las conferencias en las que ha participado Marcela Vegetti:

```
{{#ask:[[Category:Icom conf Conference]]
[[icom core participant::MarcelaVegetti]]
    |?Icom core organizer
    |?Icom core Description
    |format=table }}
```

Esta consulta, podría ser incluida en el template ***icom core Person*** de manera que para cada página de esta categoría se muestre la lista de conferencias de las que participó. Para ello es necesario cambiar el nombre de la persona por la variable {{PAGENAME}} que almacena el nombre de la página donde la consulta está incluida. Sin embargo, también es posible ejecutarla desde [Special:Ask]. Esta página, cuya funcionalidad está implementada por la extensión SMW, permite realizar consultas definidas por los usuarios, más allá de las incluidas en las páginas wiki. En la Fig. 6 se muestra los resultados obtenidos al ejecutar la consulta mostrada desde esta página especial.



Previous **Results 1– 14** Next (20 | 50 | 100 | 250 | 500)

	Icom core organizer
VirtualCall 2012 12 13	OntologySummit2013 Space prototype
VirtualCall 2013 01 24	OntologySummit2013 TrackB Space prototype
VirtualCall 2013 01 31	OntologySummit2013 TrackA Space prototype
VirtualCall 2013 02 14	OntologySummit2013 TrackD Space prototype
VirtualCall 2013 02 21	
VirtualCall 2013 02 7	OntologySummit2013 TrackC Space prototype
VirtualCall 2013 03 28	OntologySummit2013 Space prototype
VirtualCall 2013 04 11	OntologySummit2013 Space prototype
VirtualCall 2013 04 18	OntologySummit2013 Space prototype
VirtualCall 2013 04 25	OntologySummit2013 Space prototype
VirtualCall2013 02 21	OntologySummit2013 Space prototype
VirtualConference 2013 03 07	OntologySummit2013 TrackA Space prototype
VirtualConference 2013 03 14	OntologySummit2013 TrackC Space prototype
VirtualConference 2013 03 21	OntologySummit2013 TrackD Space prototype

Previous **Results 1– 14** Next (20 | 50 | 100 | 250 | 500)

Fig. 6. Resultados de la consulta acerca de conferencias a las que asistió una persona dada

Es importante resaltar, que una vez creados los datos en una wikimedia semántica, éstos no necesitan mantenerse dentro de la wiki; pueden ser fácilmente exportados a otros formatos como archivos separados por comas, JSON y RDF. Esto posibilita que una wiki semántica sirva como fuente de datos de otras aplicaciones. En particular, utilizando la página especial ontolog-02. cim3.net/wiki/Special:ExportRDF es posible

exportar los datos de una o varias página como tripletas RDF, con lo cual el contenido de la cumbre de ontologías 2013 es ahora legible y descubrible por máquinas. Impulsando con esto el movimento del contenido de ontolog PSMW hacia la obtención de las 5 estrellas que según Berners-Lee deben tener los datos para pertenecer a LinkedData [4].

4 Conclusiones y Trabajos Futuros

Ontolog es una comunidad abierta virtual que busca impulsar avances en el campo de la ingeniería ontológica y las tecnologías semánticas. Todas las contribuciones realizadas por los miembros de esta comunidad en las diferentes actividades que se proponen son almacenadas en una plataforma wiki que no tiene implementadas ninguna de las tecnologías semánticas que este foro trata de promover.

Otra de las actividades que se llevan a cabo en el marco de Ontolog, es el desarrollo de la plataforma PSMW, a la cual se está migrando el repositorio de Ontolog. El cambio de plataforma, sólo provee la capacidad de añadir semántica, lo cual debe ser complementado con la implementación de una ontología en la plataforma PSMW, para incorporar efectivamente semántica. En este artículo se propone la utilización de la ontología ICOM para representar la estructura y el contenido de la cumbre de ontologías 2013 y capturar el contenido de acuerdo a este vocabulario común. Adicionalmente, se propone la reorganización del contenido de las cumbre de manera que fomente y facilite el acceso y la reutilización del material generado.

La implementación se está llevando a cabo a nivel de prototipo. Una vez validados los resultados de esta propuesta, se procederá a extender la aplicación de la ontología para la representación de otras actividades de Ontolog.

Agradecimientos

Este trabajo ha sido financiado en forma conjunta por CONICET, la ANPCyT (PAE-PICT2315 y PIP 112-200801-02754) y la Universidad Tecnológica Nacional (PID 25/O156 y PID 25/O144). La comunidad Ontolog ha brindado su infraestructura para permitir la realización de este trabajo. Se agradece el apoyo brindado por estas instituciones, así como la colaboración de Tejas Parikh, Ken Baclawski, Ali Hashemi, Peter Yim y Soledad Sonzini para llevar adelante esta propuesta.

Referencias

[1] Krötzsch, M., D. Vrandečić, M. Völkel, H. Haller, R. Studer (2007). Semantic Wikipedia. Web Semantics,5,

[2] Baclawski,K., V.Gupta, T. Parikh, P. Yim, J. Cheyer. (2008). Purple MediaWiki: Fine-Grained Addressability of Wiki Content. Available at: http://project.cim3.net/wiki/PMWX_White_Paper_2008

[3] OASIS ICOM. (2013). Integrated Collaboration Object Model (ICOM) for Interoperable Collaboration Services Version 1.0. 31. OASIS Committee Specification 01. Disponible on-line en: http://docs.oasis-open.org/icom/icom-ics/v1.0/cs01/icom-ics-v1.0-cs01.html

[4] Tim Berners-Lee. (2009). Linked Data. Disponible en línea en: http://www.w3.org/DesignIssues/LinkedData.html

Applying EDON Methodology and SBVR2OWL Mappings for Building an Ontology-Aware Software

Cecilia Gaspoz, Valeria Bertossi, Emiliano Reynares,Ma. Laura Caliusco

CIDISIResearch Center – UTN – FRSF – Lavaise 610 – Santa Fe – Argentina
ceciliagaspoz@gmail.com, valeriabertossi@live.com.ar,
{reynares,mcaliusc}@frsf.utn.edu.ar

Abstract. In Ontology-Aware Software, ontologies are use at run time to, for example, use their content in operations of information searching or as database substitutes for information storage. In order to integrate the software development and ontology building processes, involved in building ontology-aware information system a methodology called EDONhave defined. The main disadvantage of this methodology is that the heuristic to generate an implemented ontology from the requirement elicitation is not complete enough. On the other hand, recently, the Object Management Group (OMG) has standardized a language called Semantics of Business Vocabulary and Rules (SBVR) and different approaches have been proposed to map SBVR expressions into the OWL ontology language. In this paper, we report our experience in developing an ontology-aware information system by using an adaptation of the EDON methodology including the SBVR2OWL mappings.

Keywords: Ontology-Aware Information System, SBVR2OWL Mappings

1 Introduction

Since the latter part of the 20th century there has been a growing interest in applying the ontology in the context of software engineering due to the advent of the Semantic Web and the technologies for its realization. In the software engineering context, an ontology can be used at run time in two different ways: (1) as Architectural Artifacts (Ontology-Driven Software), ontologies are used as central elements of the proposed software architecture, and (2) as Information Resources (Ontology-Aware Software), ontologies are used at run time in order to, for example, use their content in operations of information searching or as database substitutes, for information storage [3].

In the context of ontology-aware software, developers have to face the problem of how to integrate software development and ontology building methodologies assuring the project success. The development of methodological approaches for building an ontology as software artifact is still an open research area. There are many languages, techniques and tools for the representation, design and construction of ontologies [5]. But the great majority of these have been created for and by the knowledge

engineering community. Because of this, the use of ontologies by Software Engineering professionals and researchers can be seen as an additional learning experience, and in some cases, of considerably great effort [3]. Moreover, a survey [14] showed that approximately 50 % of its participants did not use any ontology engineering methodology in large-scale projects.

In order to avoid this problem, Reynareset. al. [13]have defined a methodology called EDON to build an ontology-aware system. This methodology proposes to develop an ontology that fulfills the requirements of the development cycle to which it belongs. From requirements, through CQs and LELs, you get the necessary information about the domain which is then captured as objects, relationships and properties in the implemented ontology. With regard to CQs, they can lead to create objects, relations or properties that are not relevant to the system, but they are for the environment in which the system is embedded. This happened to us in our development and is mainly due to those who are not familiar with the development of ontologies think in terms of the system. With regards to LELs, the heuristics used to build the ontology from them is not complete enough [1]. Then, although the ontology conceptualization by using CQ and LELs has proven to be useful to facilitate the communication among the DEs, SEs and KEs, a more powerful formalism will improve the way complex business rules are expressed.

Recently, the Object Management Group (OMG) has standardized another language called Semantics of Business Vocabulary and Rules (SBVR) [10]. SBVR has been conceptualized for business people and designed to be used for business purposes independent of information systems designs. The linguistic approach adopted by the proposal enables the expression of business knowledge through statements rather than diagrams. That is rooted in the insight that diagrams are helpful for depicting structural organization of concepts but they are impractical as a primary means of defining vocabularies and expressing business rules. Different approaches have been proposed to map SBVR expressions into OWL language [12].

The objective of this paper is to report our experience in developing an ontology-aware information system by using an adaptation of the EDON methodology including the SBVR2OWL mappings defined by Reynares et al. [12]. To this aim, the paper is organized as follows. Section 2 defines the concepts necessary to understand the content of this paper. Section 3 describes the development the Ontology-Aware Information System. Finally, Section 4 is devoted to discussion and lessons learned.

2 Conceptual Foundations

2.1 Evolutionary Development of ONtologies (EDON)

EDON [13] is an approach for building from scratch an ontology intended to be used as a structural conceptual model of an information system, encoding business rules in a declarative way. EDON adopts a requirement driven, iterative, and incremental approach and it is composed by the processes described next.

Requirements Selection Process. This process is composed by three activities: (1) identification of the functional requirements that involves business rules in their meeting, (2) identification and prioritization of the domain entities involved in the meeting of the requirements identified before, and (3) requirements grouping and selection according to the importance of the entities involved.

Ontology Development Process. This process involves Development Activities that allows evolving from an abstract model toward a computable ontology, and Support Activities are carried out along the whole development process. The Development Activities are: specification, conceptualization, formalization, refinement, implementation and alignment. The Support Activities are: knowledge elicitation and evaluation. The techniques to carry out them are based on the different methodologies and good practices for building ontologies developed since mid-1990 [5]. EDON considers the performing of the refinement activity with the aim of extending the ontology by focusing on the declarative formulation of business rules.

Ontology Alignment Process. Each application of EDON produces an ontology that supports a disjoint set of functional requirements, i.e., those selected on the specification activity of the iteration. Therefore, the alignment of current and previous version of the ontology is needed as a way to support both set of requirements. Ontology alignment is the process of determining the different types of (interontology) relationships among their terms [11]. As a result, a new ontology composed by sub-ontologies is created.

2.2 SBVR2OWL Mappings.

SBVR.SBVR [10] defines the vocabulary and rules for documenting the semantics of business vocabularies, business facts, and business rules; which allows their verbalization in a controlled vocabulary readily understandable by business people. The fact-oriented approach of SBVR stems from the Business Rules Manifesto [2], stating that rules builds on facts, and facts build on concepts as expressed by terms. Therefore, terms express business concepts, facts make assertions about these concepts, and rules constrain and support these facts. SBVR supports such approach by providing noun concepts and verb concepts respectively corresponding to the notions of terms and facts.

As early stated, SBVR adopts a linguistic approach that allows to define vocabularies and express operative rules. According to this insight, SBVR defines a Controlled Natural Language (CNL) and describes the way to mechanically mapping such CNL expressions to SBVR formal concepts.

OWL. The OWL 2 Web Ontology Language (OWL 2) is the latest version of an ontology language proposed by the World Wide Web Consortium (W3C) [16]. OWL 2 ontologies provide classes, properties, individuals, and data values, and are stored as Semantic Web documents. An OWL 2 ontology is a formal description of a domain of interest rooted in three syntactic categories that are interpreted under a standardized semantics, which allows useful inferences to be drawn.

- Entities, such as classes, properties, and individuals. They are the basic elements of an ontology and are identified by Internationalized Resource Identifiers (IRIs) [7].
- Expressions, representing complex notions in the domain being described.
- Axioms, which are statements asserted to be true in the domain being described.

OWL 2 ontology language defines several concrete syntaxes that can be used to serialize and exchange ontologies. Among them, the functional style syntax is defined in the OWL 2 structural specification [7] with the aim to state the semantics of OWL 2 constructors and allow a compact writing of ontologies.

Mappings. Mappings defined by Reynareset. al. [12] allow the automatable generation of an OWL2 ontology from the SBVR specifications of a business domain. Transformations are rooted on the structural specification of both standards and are depicted in subsections below by grouping and sequencing them according to the inherent logical order of the subject matter itself. In addition to their theoretical expression, the mappings are illustrated by building an ontology that reflects the business knowledge exposed by a case study. Some of these mappings are shown in Table 1.

Table 1. An excerpt of the Mappings defined by Reynares et.al.[12]

1. Each object type ot is mapped to Declaration(Class(a:ot))
2. exactly-n Quantification, where "n" is a non-negative integer:
– If the logical formulation scopes over a unary fact type, the expression is mapped to DataExactCardinality(n a:DataPropertyOne a:DataRangeOne)
– If the logical formulation scopes over a binary fact type, the expression is mapped to ObjectExactCardinality(n a:ObjectPropertyOne a:ClassOne)

3 Applying EDON and SBVR to OWL2 Mappings for developing an ontology-based system.

The methodology applied in the development of the fellow recommender system is based on EDON methodology[13], which was adapted, in the experience describe in the following subsections, to include the SBVR to OWL Mappings[12].

3.1 Requirements Selection.

Requirements were classified in two classes: those requirements that will be supported by the ontology and those which will not. Some requirements of the first class were selected to implement in a first iteration of the development process. A storyboard exposing a functional requirement belonging to the selected subset is: *"The*

system should evaluate the indicators involved in the point assignment process for each one of the candidates and the order of those candidates based the general indicator. The ranking should be display on screen."

Alumno (Student), Materia (Subject), Universidad (University), Facultad, Carrera (Career), PlanDesarrolloAcadémico (AcademicDevelopmentPlan), Beca (Fellowship), SituaciónAcadémica (AcademicSituation), SituaciónEconómica (EconomicSituation) were identified as the core entities involved in the meeting of the requirements identified before.

3.2 OntologyDevelopment.

Specification. Based on the core entities identified before and the general knowledge of the problem, Competency Questions (CQs) were proposed. An excerpt of them is showed in Table 2. From the CQs, a list of the domain entities needed for answering them was identified.Some of these domain entities are:Postulante (Applicant - student enrolled in a fellowship), Candidato (Candidate – applicant who meets every requirement), Becario (Fellow – Candidate to whom the fellowship has been granted).

Table 2. An excerpt of the Competency Questions

– Are every applicant to the university fellowship registered during the registration period?
– Are all candidates regular students?
– Which are the aspects related with the academic situation of the candidate that impact in the ranking process?
– Is the list of the candidates order by decreasingly based in the general indicator?
– Has every candidate approved at least 5 subjects during the last school year? Those who not, ¿are those new students?

Conceptualization. In this activity, the knowledge about the domain entities was collected from the information sources: the university's fellowship regulations and a fellowship management report. The business rules extracted from these resources were written in natural language, in order to represent them independently of the modeling paradigm and the implementation language of the target ontology.

Formalization. The business rules identified were translated from the natural language to SBVR. This activity includes: Recognize the noun concepts, the fact types and keywords; differentiate noun concepts belonging to complex concepts from noun concepts belonging to datatypes, re-elaborate the fact type according to the fact being represented, build the business rules by applying restrictions on the statements.

Then, the business vocabulary was organized by means of vocabulary entries, as shown in Table 3.

Table 3.SBVR specification of "Beca" (Fellowship) concept.

Beca
• Definitions: • General Concept: • Concept Type: Object Type • Necesity: – each becatieneciclo exactly one ciclobeca – eachbeca tiene plazo inscripción exactlyoneplazo inscripción – eachbeca tiene valor canasta familiar exactlyonecanasta familiar • Posibility:
Ref: tieneciclo: has school year - ciclobeca: school year of the fellowship - plazo de inscripción: registration period - canasta familiar: basic market basket indicator.

Ontology Implementation. In order to create the ontology implementation, the SBVR2OWL transformations, defined by Reynareset. al.[12], were apply to the SBVR model of the business vocabulary created in the previous activity. An example of this process is showed in Table 4. The ontology was implemented using the free ontology editor called Protégé and the Pellet inference engine that provides sound-and-complete OWL-DL reasoning services. The ontology was written in OWL-DL 2.0 ontology language and serialized in OWL/RDF format.

Table 4. OWL specification of "Beca" concept.

```
Declaration(Class(BecaUTN:Beca))
SubClassOf(BecaUTN:Beca
    ObjectMinCardinality(1 BecaUTN:tieneCicloBecaUTN:CicloBeca)))
SubClassOf(BecaUTN:BecaObjectMinCardinality(1
    BecaUTN:tienePlazoInscripcionBecaUTN:PlazoInscripcion)))
SubClassOf(BecaUTN:Beca
    DataExactCardinality(1 BecaUTN:CanastaFamiliarxsd:float))
```

Refinement. The resulting ontology represents the main concepts of the problem domain. The refinement activity consists in further extending the ontology by focusing on the formulation of rules, which are obtained from the knowledge and information sources identified in the specification activity. The rules allow implementing the algorithm for making the fellows' ranking, and several classifications, e.g. each instance of Alumno (Student) can be classified in Postulante (Applicant) and/or Candidato (Candidate); each instance of Examen (Test) is classified in ExamenAprobado (ApprovedTest) and ExamenNoAprobado (FailedTest), etc.

The rules were implemented in the Semantic Web Rule Language (SWRL), which provides the ability to express Horn-like rules in terms of OWL concepts [9]. Table 5 shows some of the rules implemented in the study case.

Table 5. An excerpt of the rules implemented in SWRL

– Examen (?examen), calificacionExamen(?examen,?calificacion), greaterOrEqual(?calificacion, "4"^^UnsignedShort) →ExamenAprobado(?examen)
– Examen (?examen), calificacionExamen(?examen,?calificacion), lessThan(?calificacion, "4"^^UnsignedShort) →ExamenNoAprobado(?examen)
– Alumno(?alumno), esIngresante(?alumno, true) →AlumnoRegular(?Alumno)
– Alumno(?alumno), esIngresante(?alumno, false), rinde(?alumno,?exam1), rinde(?alumno,?exam2), ExamenAprobadoCicloAnterior(?exam1), ExamenAprobadoCicloAnterior(?exam2), DifferentFron(?exam1, ?exam2), →AlumnoRegular(?Alumno)

3.3 Ontology Evaluation.

Quality evaluation task was performed by means of OQuaRE [4], a framework conceived for that purpose and based on the SQuaRE standard for software quality evaluation [6]. OQuaRe defines a quality model which is divided into a series of characteristics organized into subcharacteristics which are evaluated by applying a set of automatable metrics. OQuaRE defines the criteria to transform the quantitative scores of each metric into a 1-5 range and establishes that 1 means not acceptable, 3 is minimally acceptable and 5 exceeds the requirements. After such transformation, score for each subcharacteristic is the mean of its associated metrics while the score of each characteristic is the mean of its sub-characteristics. The set of characteristics scores is the quality assessment result, enabling the identification of strengths and flaws of the ontologies rather than simply pointing out a "best ontology". Dimensions evaluated, shown in Figure 1, are defined as follows:

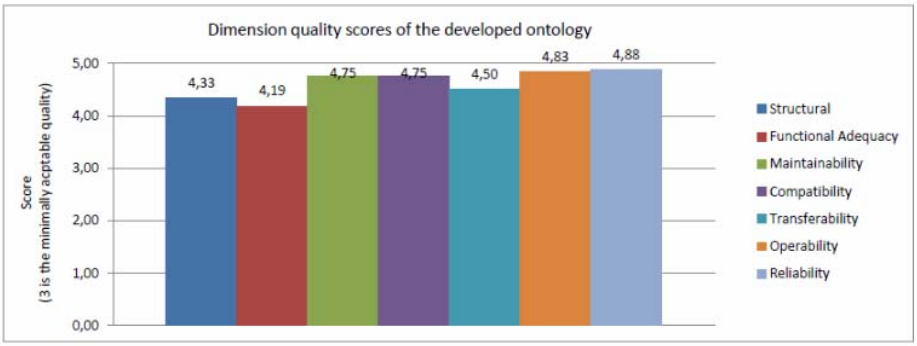


Fig.1. Characteristics scores of the ontology developed

- Structural dimension involves formal and semantic properties that are important when evaluating ontologies since it accounts for quality factors such as consistency, formalization, redundancy or tangledness.
- Functional adequacy dimension refers to the appropriateness of the ontology for its intended purpose, according to the categories identified by [15].

- Maintainability dimension is related to the capability of the ontologies to be modified for changes in the environment, in requirements or in functional specifications.
- Compatibility dimension refers to the ability of two or more ontologies to exchange information and/or to perform their required functions while sharing the same hardware or software environments. The compatibility dimension can be evaluated over a single ontology - although intuitively it involves properties about more than one ontology - given that it is quantitatively assessed by means of a set of metrics applied to each ontology separately.
- Transferability dimension is the degree to which the ontology can be transferred from one environment (e.g., operating system) to another.
- Operability dimension refers to the effort needed to use the ontology and, in the individual assessment of such use, by a stated or implied set of users.
- Reliability dimension is the capability of the ontology to maintain its level of performance under stated conditions for a given period of time.

A quickly recognizable outcome is the level of quality shown by the ontology: according to the meaning assigned for OQuaRE to the values of the 1-5 ranking system, it largely outperform the minimally acceptable quality in all considered dimensions. Moreover, the global quality score - which is equal to 4.60 and it is calculated as the mean of all the scores - is very close to the maximal quality value.

3.4 Ontology Alignment.

The alignment activity consists in determining the different types of (inter-ontology) relationships among their terms [8] [15].As a result, a new ontology composed by sub-ontologies is created. The first version of the ontology does not involve the performing of alignment activities. As single iteration of this EDON adaptation was performed, this activity was not required.

3.5 FellowRecommender System Implementation.

After the ontology evaluation, the Fellow Recommender System was implemented in Java by using the JENA framework. This Software includes inscription, academic plan's punctuation and fellow's ranking functions, as shown in Figure 2.With regards the software quality, the functionality, efficiency, reliability and maintainability are closed related with those measured by the ontology since it is the core of the system. The usability was evaluated by a domain expert who gives a useful feedback to improve our system in a further work.

4 Discussion And Lessons Learned

In this paper we have reported our experience and showed the satisfactory results in developing an ontology-aware Fellow Recommender Systems using the EDON Method adapted to includeSBVR language to write business rules, in the formalization activity of the ontology development process, and SBVR to OWL2 Mappings, to be used during implementation activity.

Business Rules are usually embedded in the procedural part of the application. Using Ontologies to encapsulate them, made easier the modification and adaptation processes, allowing to use these system in others environments, such as other universities, without making a lot of changes. This is because, in order to adapt the system to other universities, the set of Business Rules defined in the ontology, is the only thing that have to be modified.

Related with EDON some advantages can be mentioned that we identified form this experience. The use of CQs can lead you to identify objects, relations or properties, that are part of the domain of the problem that is attach by using an ontology-aware system, as well as restrictions, which can guide you in the ontology testing process.

On the other hand EDON proposed to align the ontologies that are developed throughout the history of the system, allowing and providing the system to grow. This is an important feature to get extensible systems, which could adapt to new requirements, e.g. handling a new type of fellowship.

Finally, Adding SBVR and SBVR2OWL Mappings to EDON Methodology, made the ontology development process of this system easier and fluid, making the transition from the business rules to the implemented ontology, natural, simple and intuitive, focusing in the conceptualization and formalization process and without taking great efforts during the implementation process.

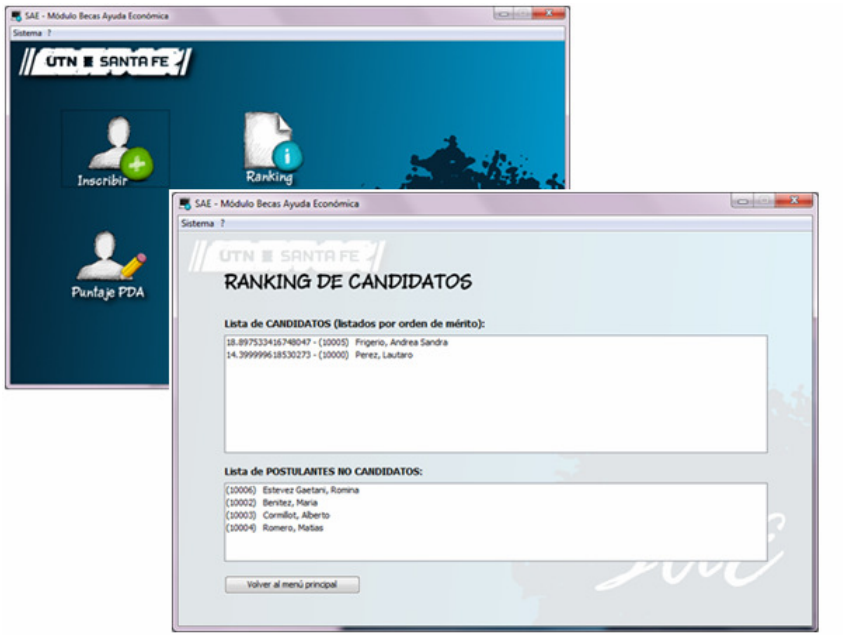


Fig. 2. The ontology – aware system implemented

5 References.

1. Ayub, C.; Cian, A.;Reynares, E., Caliusco, Ma. L. "An Experience Report on using the EDON Method for Building a Team Recommender System". 42° Jornadas Argentinas de Informática - Simposio Argentino de Ingeniería de Software. Córdoba (Córdoba - Argentina) Septiembre 2013. En prensa.
2. Business Rules Group: Business Rules Manifesto. Version 2.0. Accessible in: http://www.businessrulesgroup.org/brmanifesto.htm (2003)
3. Calero, C. and Ruiz, F. and Piattini, M., eds.: Ontologies for Software Engineering and Software Technology. Springer. (2006)
4. Duque-Ramos, A. and Lopez, U. and Fernandez-Breis, J. T. and Stevens, R. and Aussenac-Gilles, N.: OQuaRE: a SQuaRE-based approach for evaluating the quality of ontologies. Journal of Research and Practice in Information Technology, 43, 159- 173.
5. Gómez-Pérez, A. and Fernández-López, M. and Corcho, O.: Ontological Engineering. Springer/Heidelberg. (2004).
6. International Organization for Standardization (ISO) ISO/IEC 25000 2500, Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE)
7. Internet Engineering Task Force (IETF): RFC 3987: Internationalized Resource Identifiers (IRIs). Accessible in: http://www.ietf.org/rfc/rfc3987.txt (2005)
8. Liu, P. and Dew, P.: Using semantic web technologies to improve expertise matching within academia. In Proc.: I-KNOW '04, pages 370–378, (2004)
9. O'Connor, M. Knublauch, H., Tu, S. &Musen M.: Writing rules for the semantic web using SWRL and Jess. In: Proc. 8th Int. Protégé Conference, Protégé with rules, (2005).

10. Object Management Group (OMG): Semantics of Business Vocabulary and Business Rules (SBVR). Version 1.0: Formal Specification. Accessible in: http://www.omg.org/spec/SBVR/1.0/ (2008)
11. Pavel, S and Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. In IEEE Transactions on Knowledge and Data Engineering, PP-99 (2011).
12. Reynares, E. and Caliusco M. A. and Galli, M. R.: An Automatable Approach for SBVR to OWL2 Mappings. In Proc. XVI Ibero-American Conference on Software Engineering (CIbSE 2013) Montevideo, Uruguay. (2013)
13. Reynares, E. and Caliusco M. A. and Galli, M. R.: EDON: A Method for Building an Ontology as Software Artefact. In Proc. 41st Argentine Conference on Informatics - 13th Argentine Symposium on Software Engineering (JAIIO - ASSE 2012) La Plata, Buenos Aires, Argentina. (2012)
14. Simperl, E., Mochol, M, and Bürger, T: Achieving Maturity: the State of Practice in Ontology Engineering in 2009. International Journal of Computer Science and Applications, 7(1):45 – 65, 2010.
15. Stevens, R. and Wroe, C. and Gobel, C. and Lord, P.: Application of ontologies in bioinformatics. In Handbook of Ontologies in Informations Systems. Staab, S. and Studer, R. (eds). pp 635-658. Springer. (2008).
16. World Wide Web Consortium (W3C): OWL 2 Web Ontology Language. Structural Specification and Functional-Style Syntax. Accessible in: http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/ (2009).

Propuesta de tecnología móvil para la administración de información vinculada a la gestión de espacios áulicos

Martín S. Martinez, Sonia I. Mariño, Pedro L. Alfonzo, Maria V. Godoy
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste. Corrientes. Argentina.
seba_martinezz@hotmail.com; plalfonzo@hotmail.com, simarinio@yahoo.com

Abstract. El avance de la informatización en las actividades humanas es un hecho, en las últimas décadas se ha incrementado la cantidad de software producido reflejándose en cambios en diversos aspectos de la sociedad del conocimiento. El ámbito educativo no escapa a esta realidad. Por lo expuesto, en este trabajo se presenta una segunda versión de un sistema web para la administración de espacios áulicos; incorporándose como tecnología emergente el acceso a información desde dispositivos móviles. El prototipo fue modelizado a partir de la gestión de espacios físicos de la Facultad de Ciencias Exactas y Naturales y Agrimensura (FaCENA) – Sede 9 de Julio, de la Universidad Nacional del Nordeste (UNNE), con miras a adecuarse a otras instituciones de Educación Superior.

Keywords: Educación Superior, Ingeniería web, sistemas de gestión web, administración automatizada de espacios físicos.

1 Introducción

En los últimos años el crecimiento exponencial de la tecnología ha producido grandes avances en la informatización de las actividades. El bajo costo es cada vez mayor, los productos en la actualidad se encuentran muy cercanos al alcance de todos y las acciones de alfabetización e informatización están en ascenso. Lo expuesto se refleja en una sociedad cada vez más informatizada que requiere al mismo tiempo de diversas soluciones de software para el máximo aprovechamiento de la información.

Se coincide con [1] en que el surgimiento de la cuarta generación (4G) de celulares, conocidos como teléfonos inteligentes (smartphone), permitió una revolución en el desarrollo de software. La demanda está en constante aumento (Fig. 1). Años atrás lo más común era el desarrollo para máquinas de escritorio, el avance de la tecnología permitió crear dispositivos potentes y del tamaño para caber en la palma de una mano.

En los ámbitos de la educación superior, surgen diversos requerimientos tecnológicos, la mayor orientación ha sido informatizar el proceso de enseñanza y aprendizaje mediante la introducción de modalidades identificadas como *e-learning*, *b-learning* y *m-learning*, entre otras. Además, en la administración y gestión de los procesos desarrollados en el sector educativo existen numerosas soluciones mediadas

por las Tecnologías de la Información y Comunicación, Sin embargo aquellas orientadas a gestionar la asignación de espacios físicos como son los laboratorios, salas de conferencias y aulas, rige generalmente bajo la responsabilidad de los "bedeles", quienes administran los espacios contemplando datos referentes a horarios de inicio y finalización de clases o eventos académicos, profesor designado, materia o evento asignado, entre otras variables.

En [2], se detalla el desarrollo del sistema informático bajo un enfoque para consultas de "escritorio", orientado a computadoras de tipo Notebook, Netbooks, PC. A partir de una etapa de validación con los potenciales usuarios "bedeles", quienes utilizaron el software y con miras a mejorar o incorporar otras funcionalidades a partir de información de realimentación obtenida, en este trabajo se describe la incorporación de una tecnología emergente como es la consulta desde dispositivos móviles.

Cabe aclarar que el sistema informático se modeló considerando como contexto de implementación la Facultad de Ciencias Exactas y Naturales y Agrimensura (FaCENA), sede 9 de Julio, de la Universidad Nacional del Nordeste (UNNE), situada en Corrientes, Argentina. En su especificación primó proporcionar una solución fiable y acorde con las necesidades y requerimientos edilicios demandados por las distintas carreras dictadas en el mismo.

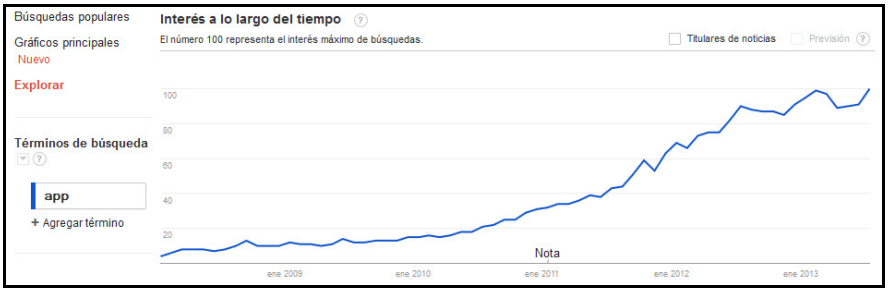


Fig. 1. Crecimiento de la demanda de aplicaciones móviles – Argentina. Fuente: Google Trends.

2 Metodología

En [2] se detalló la metodología aplicada en la construcción del sistema informático para la gestión de espacios físicos, su desarrollo se fundamentó en el modelo de prototipos incrementales o evolutivos como ciclos de vida [3], [4]. [5], [6]. Es así como se consideró que al comienzo del proyecto, hay partes del sistema que no están del todo claras y generalmente el usuario final no especifica todos los requerimientos al inicio del mismo, completándose a fin de ajustar a lo largo del diseño y desarrollo. A continuación se describen las etapas abordadas en esta segunda versión:

ETAPA 1. La presentación de una versión preliminar descripta en [2] constituyó un medio de obtener datos para refinar el sistema, en este caso se detectó la necesidad del acceso a consultas desde dispositivos móviles.

- Análisis de un sistema informático para la gestión de espacios físicos. Se revisó el diseño preliminar, desarrollándose una segunda versión de prototipo a fin de integrar tecnología móvil. Cabe aclarar que se continuó aplicando modelado UML (Unified Markup Language) para redefinir las funciones a partir de los nuevos requerimientos utilizándose: i) Casos de usos ii) Conversaciones de los casos de uso para comprender que debería hacer cada uno, y iii) Diagramas de secuencia para identificar los diferentes flujos de información necesarios.
- Perfiles de usuarios. En cuanto al acceso de la información se aclara que se continuó trabajando con los dos perfiles mencionados en [2], "invitados" y "usuarios". El último está compuesto por: bedeles y administrador.

ETAPA 2. Diseño del sistema informático, en esta segunda versión del prototipo, se abordaron las siguientes actividades:

- Diseño de interfaces. Orientadas a las nuevas funcionalidades incorporadas, considerando los diversos perfiles de usuarios previstos.
- Diseño de la base de datos. Se rediseñó la base de datos y sus posibles relaciones con otras fuentes de datos (Fig. 2).

ETAPA 3. Desarrollo del sistema información.

- Selección de herramientas: para el desarrollo de esta nueva versión se continuó utilizando HTML (Hyper Text Markup Language) [7]. CSS - Hojas de Estilo en Cascada [8]. jQuery [9]. JavaScript [10]. PHP (Hypertext Preprocessor) [11]. WampServer (acrónimo formado por Windows, Apache, MySQL y PHP) [12]. DOMPDF [13]. MySQL Workbench [14]. Notepad++ [15]. En la generación de la solución móvil se optó por Phonegap [16] y Android Studio [17].
- Codificación de un nuevo módulo, para establecer el enlace entre la aplicación nativa de los móviles y la base de datos de sistema. Inicialmente se desarrolló para smartphones y tablets.

ETAPA 4. Puesta en funcionamiento o implementación. Las pruebas con los potenciales usuarios y destinatarios de la solución en desarrollo, brindan información de realimentación a fin de asegurar su implementación.

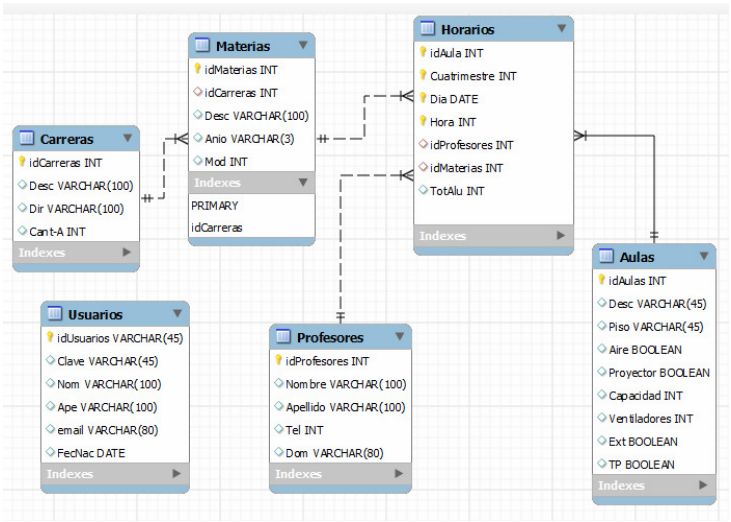


Fig. 2. Estructura de la Base de Datos y sus relaciones.

3 Resultados

En esta sección, se describe una segunda versión de un sistema informático parametrizable, orientado a la gestión de espacios físicos bajo el enfoque de soluciones móviles.

En la Fig. 3 se detalla la arquitectura de esta versión del sistema informático diseñado para la gestión de espacios áulicos. Donde las líneas azules indican las solicitudes o peticiones al servicio y las líneas rojas las consultas generadas y entregadas a las interfaces accesibles por los usuarios.

Como se especificó en [2], para acceder al sistema, se debe ingresar desde algún explorador Web. Los diferentes módulos que lo componen están dispuestos mediante la barra de "menú".

Se coincide con [18] en que el "*m-learning* promueve experiencias contextuadas y colaborativas" en estos tiempos es recomendable incluir soluciones móviles desde los espacios de Educación Superior siendo tanto las orientadas a los procesos de enseñanza y aprendizaje como también las de índole administrativa, este ultimo tipo abordada en el presente trabajo. Por ello, como tecnología emergente incorporada se resume una aplicación para la consulta de aulas mediante tecnología móvil accesible desde smartphones y tablets. En una primera etapa se destinará la solución a aquellos usuarios con acceso a la red mediante conexión Wifi, 2G, 3G y 4G. Con soporte para interfaz web y Java. Esta aplicación permite la consulta de "bolsillo" para visualizar información de horarios específicos mediante el acceso a la base de datos del sistema. (Fig. 4). En la Fig. 5 se ilustra una consulta desde la web y la misma desde una Tablet. Se proyecta, para próximas versiones de la aplicación desarrollar las consultas

no siempre de forma remota, sino de prever una sincronización diaria o semanal, para realizarlas de modo “offline”. Cabe aclarar que temporalmente, la segunda versión del sistema que se describe sólo se aplica a consultas, es decir, no es viable establecer una administración de los datos desde dispositivo móvil.

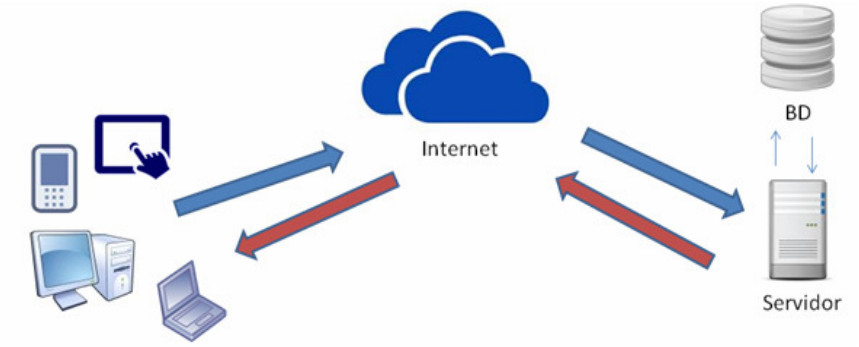


Fig. 3. Arquitectura del sistema informático de gestión de aulas

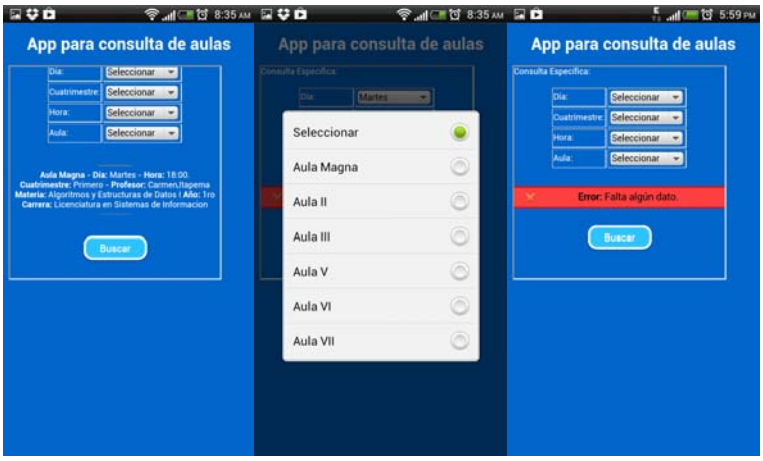


Fig. 4. Aplicación móvil para consultas.

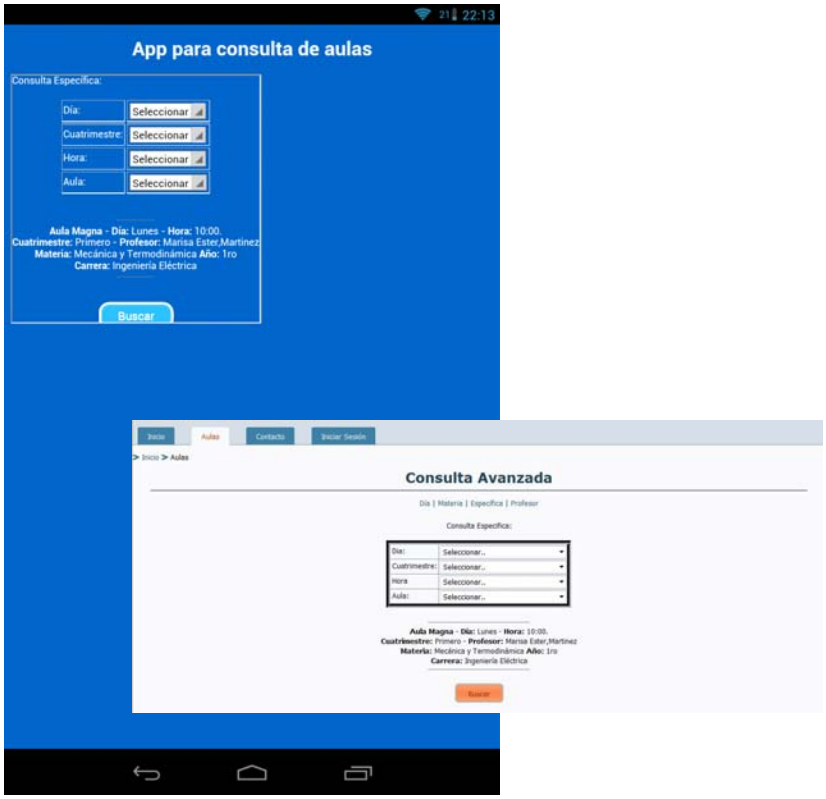


Fig. 5. Consulta desde la web y desde una Tablet

4 Conclusión

A partir de una solución tecnológica previamente desarrollada, destinada a la administración de espacios físicos de la Facultad de Ciencias Exactas y Naturales y Agrimensura – Sede 9 de Julio, de la Universidad Nacional del Nordeste (UNNE). Corrientes – Argentina diseñada para mejorar la calidad de los procesos administrativos y de logística que sirven de apoyo para las actividades de la institución, se incorporó una funcionalidad accesible desde dispositivos móviles orientada a ampliar los servicios de información sin restricciones espacio-temporales. Se prevé su conexión desde cualquier punto a través de las redes móviles como 2G, 3G, 4G y Wifi, brindando a los potenciales usuarios información siempre actualizada.

Por otra parte, se analiza la viabilidad de integrarlo al sistema SIU, tecnología informática disponible en las universidades Argentinas, y de este modo lograr interoperabilidad con otras soluciones disponibles en la unidad académica.

Agradecimiento

El tema de investigación, se encuentra incluido en una de las líneas del proyecto denominado "Sistemas de información y TIC: métodos y herramientas". El mismo fue aprobado por la Secretaria General de Ciencia y Técnica (SGCyT) de la Universidad Nacional del Nordeste, Código Nº FO13-2011, y acreditado por Resolución Nº 142/12 C.S. de la UNNE.

Referencias

1. Fantasia, J.; Ferrochio, D.; Maldonado, C.; Martinez, E.;Trujillo, H.. Desarrollo de una metodología utilizable en la construcción de tecnología móvil. VIII Workshop de Investigadores en Ciencias de la Computación (WICC), XVII Congreso Argentino de Ciencias de la Computación.Smith, (2006).
2. Martinez, M. S.; Alfonzo, P; Mariño, S. I.; Godoy, M. V. Sistema informático para la gestión de espacios físicos. Una aproximación para la FaCENA (UNNE). Revista Internacional de Tecnologia, Conocimiento y Sociedad, http://ijtes.cgpublisher.com, Aceptado para su publicación (2013).
3. Boehm, B.: A Spiral Model of Software Development and Enhancement. IEEE. 61-72pp. (1998).
4. Mariño, S.; Godoy, V.: Sistemas de información y TIC: métodos y herramientas. PI f013-11 Acreditado por la SGCyT. UNNE. Resol. 142/12
5. Pressmann R.: Ingeniería de Software: Un Enfoque Práctico. Ed. Pearson Education, S.A., Madrid. Edition 7º. (2007),
6. Sommerville, I.: Ingeniería del Software. Ed. Prentice Hall. (2005).
7. HTML, http://es.html.net
8. CSS, http://www.w3c.es
9. Jquery, http://www.javascriptya.com.ar/jquery
10. JavaScript, http://www.w3schools.com/js
11. Php, http://www.php.net
12. WampServer, http://www.wampserver.com
13. Dompdf, http://code.google.com/p/dompdf
14. Workbench, http://www.mysql.com/downloads/workbench
15. Notepad++, http://notepad-plus-plus.org
16. PhoneGap, http://www.phonegapspain.com
17. Android Studio, http://developer.android.com/sdk/installing/studio.html
18. Herrera, S I. y Fennema, M C. Tecnologías móviles aplicadas a la educación superior, IX Workshop Tecnología Informática aplicada en Educación (WTIAE), XVII Congreso Argentino de Ciencias de la Computación, (2011).

Desarrollo de aplicaciones colaborativas para Cloud Computing

María Murazzo[1*], Nelson Rodríguez[2*], Daniela Villafañe[3*], Daniel Gallardo[4#]

[1]marite@unsj-cuim.edu.ar, [2]nelson@iinfo.unsj.edu.ar, [3]villafañe.unsj@gmail.com, [4] dim_daniel87@hotmail.com

* *Docentes e Investigadores, Departamento e Instituto de Informática – FCEFyN – UNSJ*
 # Alumno Becario de la Carrera Licenciatura en Sistemas de Informacion

Abstract. En los últimos años se ha producido una masificación de las TIC (Tecnologías de la Información y las comunicaciones) como Internet, Social Medias, Cloud Computing, etc. Esto ha provocado en los usuarios un aumento de la interacción haciendo necesario contar con aplicaciones que le brinden la capacidad de intercambiar contenidos y colaborar en la realización de tareas conjuntas.

El objetivo de este trabajo es presentar una propuesta para el desarrollo de una aplicación colaborativa con herramientas propias de cloud computing y que será almacenada en el cloud y accedida mediante una interface de usuario ubicua y transparente al usuario.

Keywords: Aplicaciones Colaborativas, Cloud Computing, GAE

1 Justificación

Las nuevas tecnologías introducen diferentes formas de entender el trabajo, incrementando la colaboración del grupo para alcanzar metas comunes, tendientes a lograr una mayor productividad y rendimiento. Así, el usuario es parte de una comunidad conectada por medio de una red, ampliando los horizontes en investigación hacia la Interacción Persona – Computadora - Persona con tecnología basada en sistemas distribuidos de computación. El objetivo no es sólo mejorar la comunicación, sino generar un nuevo entorno que se comparte con otras personas pudiendo llevar a cabo actividades conjuntas bajo el paradigma de denominado Trabajo Cooperativo.

Esta colaboración, implica que el entorno de trabajo debe ser capaz de brindar a los usuarios la posibilidad de acceder a la información desde cualquier lugar, en cualquier momento y desde cualquier tipo de dispositivos. Estas características, exigen que las aplicaciones sean del tipo anywhere, cuya principal característica es la ubicuidad que les brinda a los usuarios.

En la actualidad, la ubicuidad y la colaboración solo pueden imaginarse de la mano de la omnipresencia tanto de los usuarios como de los contenidos generados por ellos.

En este sentido un concepto capaz de soportar esta omnipresencia es el Cloud Computing.

Según el NIST (National Institute of Standards and Technology), se define Cloud Computing como un modelo de servicios escalables bajo demanda para la asignación y el consumo de recursos de cómputo. Esta definición, implica ver a los recursos como infraestructura, almacenamiento, ancho de banda, etc., como una utility más capaz de ser virtualizada para permitir a los usuarios generar contenidos consumidos por otros en forma colaborativa y ubicua.

La convergencia de Internet, la Web 2.0, el Social Media, el BigData y el Cloud Computing, han generado un ámbito propicio para el desarrollo de aplicaciones que permita a los usuarios no solo interactuar con sus aplicaciones sino convertirse en un generador activo de contenidos que serán virtualizados en una plataforma agnóstica.

En función de lo analizado, el presente proyecto de beca pretende abordar la problemática del desarrollo de aplicaciones que permitan fomentar la interacción de los usuarios, mediante el intercambio de contenidos virtualizables en el Cloud Computing.

2 Introducción

En los últimos años se ha visto evolucionar tecnologías vitales para el mundo organizacional en lo que a TIC's se refiere, tales como los servicios de telefonía, las telecomunicaciones, los datacenter, etc.

Las organizaciones están preocupadas por brindar nuevos servicios reduciendo costos, Cloud Computing ofrece la posibilidad de dinamizar el abastecimiento de capacidades informáticas, en función de la demanda cambiante. Eficiencia y eficacia son conceptos que este modelo promueve a poyándose en la ubicuidad de Internet para ayudar a las empresas a extender su cobertura, llevando los recursos de TI a cualquier parte. Cloud Computing plantea un cambio de paradigma donde lo que antes era una propiedad, se convierte en un servicio, cambiando no solo la gestión de TI sino también la organización.

En este trabajo se realiza una integración entre un SaaS, un PaaS y un IaaS. El SaaS que se utiliza es el Google Apps que es una plataforma donde se realizan dominios para la empresa que lo utilizan. El SaaS Google Apps es una plataforma de comunicación y colaboración, ya que tiene múltiples funciones, no sólo proporciona correo electrónico sino que posibilita que los equipos de trabajo compartan calendarios (Google Calendar), documentación (Google Docs), o videos (YouTube) entre otros servicios (Google Sites, Gtalk, etc.). El PaaS / IaaS es Google App Engine, que es una plataforma que permite desarrollar, almacenar y ejecutar una aplicación web, en la gestión de centros de datos de Google. Todo esto programado con el lenguaje Python que permite a integración de la APIs de Google App en una aplicación desarrollada en Google App Engine [1].

3 Cloud Computing

Es un modelo que permite a las diferentes empresas adquirir el uso de servicios y la entrega de recursos, que hace referencia a estar siempre conectado en el cloud (la nube o Internet) para recibirlos. Se podría decir que esto ya venía sucediendo, pero no en cuanto a un concepto integral y definido.

Así que la pregunta es, ¿por qué no conectarse a Internet y que alguien suministre todos los servicios de computación que la organización necesita de manera simple y se facture mensualmente por ello?, de esta forma todo lo que sea computación se convierta en una utility más.

Esta idea no es nueva, se viene trabajando en este concepto desde hace algunos años, ya que es la convergencia de modelos precursores como son Utility Computing, On Demand Computing, Elastic Computing o grid computing [2].

Internet usualmente se visualiza y conceptualiza como una gran nube donde todo está conectado y donde al conectarse se suministran todos los servicios requeridos. A este esquema de trabajo se lo denomina Cloud Computing, la cual es similar a todos los esquemas antes nombrados, pero potenciada con las tecnologías de virtualización [3].

El concepto de Cloud Computing tiene como principal característica, la transformación de los modos tradicionales de cómo las organizaciones utilizan y adquieren los recursos de Tecnología de la Información (TI).

Cloud Computing, representa un nuevo tipo de valor de la computación en red. Entrega mayor eficiencia, escalabilidad masiva y más rápido y fácil desarrollo de software. Los nuevos modelos de programación y la nueva infraestructura de TI permitirán que surjan nuevos modelos de negocios.

La Cloud Computing es un modelo de aprovisionamiento de recursos TI que potencia la prestación de servicios TI y servicios de negocio, facilitando la operativa del usuario final y del prestador del servicio.

Una de las principales ventajas para las organizaciones que deciden incorporar a sus actividades servicios prestados a través de Internet es la posibilidad de reducir sus gastos de personal técnico, instalaciones, software y, sobre todo, de tareas de mantenimiento; de esta manera el retorno de la inversión es inmediato, ya que no es necesaria preinstalación ni configuración alguna.

Todo ello se realiza de manera fiable y segura, con una escalabilidad elástica, que es capaz de atender fuertes cambios en la demanda no previsibles a priori, sin que esto suponga un incremento en los costos de gestión.

La característica básica de este modelo es que los recursos y servicios informáticos, tales como infraestructura, plataforma y aplicaciones, son ofrecidos y consumidos como servicios a través de Internet sin que los usuarios tengan que tener ningún conocimiento de lo que sucede detrás.

La consultora Gartner, Inc. ha destacado las 10 principales tecnologías y tendencias que serán estratégicas para las organizaciones en 2013. Para Gartner una

tecnología es estratégica cuando tiene el potencial para un impacto significativo en la organización en los próximos tres años. Los factores que indican un impacto significativo incluyen un alto potencial para la interrupción de TI o el negocio, la necesidad de una inversión importante, o el riesgo de llegar tarde a adoptar estos cambios.

Una tecnología estratégica puede ser una tecnología existente que ha madurado y/o adquiera la aptitud de una gama más amplia de usos. También puede ser una tecnología emergente que ofrece una oportunidad para la ventaja estratégica de negocios para los primeros adoptantes o con potencial de alteración en el mercado en los próximos cinco años. Estas tecnologías tienen un impacto tangible en la organización a largo plazo, planes, programas e iniciativas [4].

Dentro de estas 10 tecnologías, Cloud Computing ocupa el tercer lugar y predice que de a poco desplazara a la PC como entorno para que los usuarios guarden su información personal. Esto, se puede corroborar con la cantidad de contenidos que se suben a Internet o más precisamente al cloud, en la actualidad. Por ejemplo, cada minuto se suben 72 minutos de videos a YouTube, se envían 100.000 mail, se envían 277.000 tweets, se procesan 2 millones de búsqueda en google, se realizan 250.000 llamadas via Skype [5].

Cloud Computing es un esquema del tipo aaS o as a Service y que a veces se expresa como XaaS o EaaS para significar Everything as a Service. En general cualquier cosa como un servicio.

Se puede dividir al Cloud Computing en las siguientes capas: *Software como Servicio (SaaS)*, *Plataforma como Servicio (PaaS)* y *Infraestructura como Servicio (IaaS)*.

De esta forma cualquier organización que desee servicios de TICs podrá implementar un esquema XaaS y eliminar todos sus requerimientos internos y contratar sus necesidades en estas áreas externamente a cambio de un pago mensual, sin inversiones de capital [6].

4 Aplicaciones Colaborativas en el Cloud

Si bien el concepto de Aplicaciones Colaborativas no es nuevo, ha convergido para fusionarse con otras tecnologías como el Cloud Computing. De esta manera el usuario podrá acceder a aplicaciones que permitan la colaboración y el intercambio de contenidos de forma transparente, sin preocuparse de la heterogeneidad de formatos y la disponibilidad de recursos.

Estas características son muy importante considerando que las exigencias y requerimientos de los usuarios tanto a nivel profesional como social han cambiado y se han ampliado. Las principales características que requieren de las aplicaciones son ubicuidad, disponibilidad, omnipresencia, localización, inmediatez y personalización debido a estas exigencias, se hace necesario depender de la cloud para la distribución de los servicios.

En este contexto, la convergencia del Cloud Computing y el Social Media ha provocado la necesidad de desarrollar aplicaciones que permitan la colaboración de usuarios los cuales trabajan en un ambiente de red distribuido. Además, será necesaria la interoperabilidad con aplicaciones de otros usuarios o con aplicaciones comerciales como GoogleDoc, Picasa, Facebook, etc., esto exige que las aplicaciones realicen el control de concurrencia y acceso a los recursos compartidos, con el objeto de salvaguardar la integridad y la consistencia de los contenidos.

En función de lo antes expresado, los usuarios necesitan aplicaciones que les permitan tener un ambiente de colaboración sin que se deban preocupar por detalles de diseño ni mantenimiento [7].

Además, y desde el punto de vista del desarrollador, esta forma de trabajo que plantea el Cloud Computing es interesante, pues tienen la posibilidad de realizar aplicaciones al estilo web service, los cuales podrán ser consumidos por cualquier otra aplicación. Estas aplicaciones podrán ser desarrolladas mediantes plataformas PaaS que permitan trabajar en un ambiente de recursos e infraestructura heterogéneos.

5 Selección de la Herramienta de desarrollo

Dada la relevancia que ha adquirido el paradigma “cloud computing” en los últimos años y el prometedor futuro que se le presupone por delante, son muchas las empresas y organizaciones que se han posicionado o intentan hacerlo sobre las demás ofreciendo este tipo de servicios. Entre ellas, destacan Google App Engine (GAE) [8], Amazon EC2 [9] y Windows Azure [10], que proveen aplicaciones comunes en línea accesibles desde un navegador web, mientras el software y los datos se almacenan en los servidores. Todas ellas se basan en el mismo paradigma, pese a que cada una posee sus particularidades y en algunos casos existen diferencias notables entre ellas.

En este trabajo, se ha decidido usar con la plataforma GAE debido a la posibilidad de manipular en forma nativa las API de Google.

6 Google App Engine (GAE)

GAE es una plataforma concebida para desarrollar, alojar y ejecutar aplicaciones web sobre la infraestructura Google. La plataforma hace uso del paradigma cloud computing, mediante la virtualización de aplicaciones a través los numerosos servidores de los centros de datos de Google, dispersos geográficamente.

La infraestructura Google es totalmente transparente para el cliente de los servicios cloud computing, quien se despreocupa de la gestión de los recursos utilizados, mientras que el usuario desarrollador por su parte es capaz de crear, mantener y actualizar sus aplicaciones. Se entiende como usuario desarrollador al programador de aplicaciones sobre la plataforma GAE, para que más tarde éstas sean ejecutadas por los clientes de los servicios. Al contrario que plataformas como Amazon EC2, que virtualizan a nivel de imágenes de máquinas virtuales, GAE ofrece su infraestructura para contener aplicaciones exclusivamente.

7 Prototipo de la Aplicación

La aplicación que se decidió construir es un ambiente colaborativo de trabajo para las cátedras Redes de la LCC y Redes y Sistemas Distribuidos de la LSI, ambas pertenecientes al Departamento de Informática de la FCEFy N de la UNSJ.

Para dicha aplicación se utilizo el framework Webapp y el lenguaje de programación Python. En la figura 1 se muestra un prototipo de la pantalla principal de la aplicación, en donde se marcan cada una de las secciones.

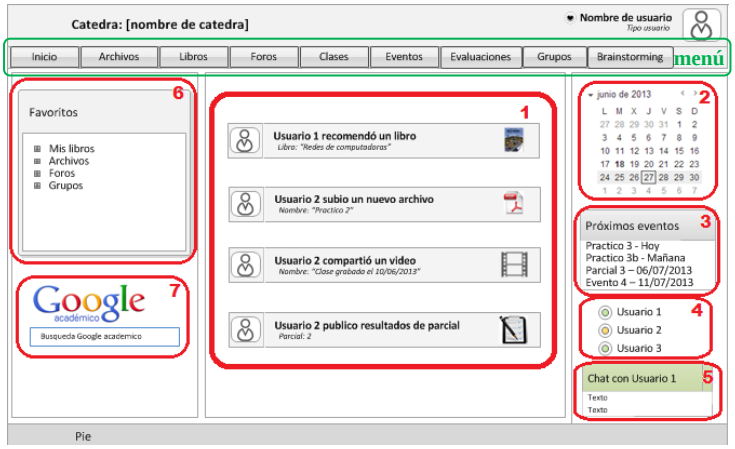


Figura 1: Pantalla principal del prototipo

A continuación se explica la funcionalidad de cada modulo:

1. Panel de novedades donde se muestran todos los eventos y actividades que realizan los usuarios, como recomendaciones, resultados de evaluaciones y publicación de archivos.
2. Calendario que muestra gráficamente cualquier evento del usuario
3. Panel que notifica los próximos 4 eventos del calendario
4. Usuarios actualmente conectados
5. Chat
6. Panel con distintos elementos favoritos del usuario
7. Búsqueda rápida en Google académico

Con respecto al menú, cada una de sus opciones se detalla a continuación:

- **Archivos**: los usuarios tendrán la posibilidad de ver, descargar y subir archivos de tipo: Audio y videos, Documentos, Planillas de cálculo, Presentaciones, Formularios, Imágenes y dibujo. Además, tanto los documentos, planillas de cálculo, presentaciones, formularios y dibujos podrán ser visualizados y editados de forma simultánea por un conjunto definido de usuarios.
- **Libros:** el usuario podrá buscar, leer y compartir libros de Google Books.

- **Foros:** contendrá un conjunto de foros categorizado por tema.
- **Clases:** se podrá acceder a un conjunto de clases presenciales grabadas en video o audio. La búsqueda se realizara por fecha.
- **Eventos:** Ver, crear y eliminar eventos para grupos de usuario tales como fechas de parciales, prácticos, resultados, etc.
- **Evaluaciones:** Se podrán crear evaluaciones escritas, prácticos, encuestas, etc para que realicen los usuarios, como así también ver los resultados obtenidos, un resumen y publicar los resultados.
- **Grupos:** Se podrán administrar grupos de trabajo, pudiendo asignar distintos privilegios, eventos y evaluaciones a cada grupo.
- **Brainstorming:** se permitirá generar distintas sesiones de brainstorming en el que podrán participar un conjunto de usuarios en tiempo real. Los resultados podrán ser guardados y vistos en cualquier momento.

Cabe destacar, la importancia del ambiente colaborativo, sobre todo en lo que ha manejo de recursos se refiere. En este caso se ha tratado de emular la forma en la que maneja el control de concurrencia DropBox, de esta manera no solo se podrán compartir documentos, sino también trabajar concurrentemente, sin preocuparse por la consistencia de los contenidos.

8 Conclusiones

Al igual que otros avances tecnológicos en el pasado, el cloud computing aporta nuevos retos y oportunidades a las organizaciones de TI y negocios. Si bien algunas de estas cuestiones son de carácter técnico (por ejemplo, rendimiento), otros son más organizacional (por ejemplo, ubicación de los datos). ¿Qué tan bien y qué tan pronto estas cuestiones se resuelven determinará si Cloud con el tiempo se consolida y puede cumplir lo que sus defensores prometen.

Este paradigma ha cambiado el centro de gravedad de la computación y tanto el ambiente académico como la industria, pero a pesar de los considerables esfuerzos e inversión existen varios problemas críticosque aun no han sido resueltos, como: portabilidad, protección de datos en ambientes Cloud, control de distribución de datos y latencia, sistemas de comunicación asíncronos en Cloud, paralelización, usabilidad de las interfaces, entre otros.

Aun a pesar de estas dificultades, el desarrollo de aplicaciones pensadas para que sean desplegadas en el cloud es auspicioso. No se puede dejar de lado la masificación de las actuales tecnologías que rodean a internet, como el cloud computing, la cual brinda la posibilidad de independizarse de la infraestructura de almacenamiento para centrarse en los contenidos, más que en el cómo hacer para que estos, estén disponible en todo momento para los usuarios.

9 Bibliografía

[1] Rodriguez, Villafañe, Murazzo, Gallardo. “GAE, una estrategia para complementar SaaS y PaaS a través de la web”. 2° SABTIC. Tres de Maio, Brasil. Agosto 2012

[2] Marston, Li, Bandyopadhyay, Zhang, Ghalsasi. “Cloud computing — The business perspective”. Decision Support Systems 51 (2011) 176-180. Elsevier. 2011.

[3] Lu, Hai-shan, Ting-ting.”Research on Hadoop Cloud Computing Model and its Applications”.2012. Third International Conference on Networking and Distributed Computing.

[4] Gartner. “Gartner Identifies the Top 10 Strategic Technology Trends for 2013”. URL: http://www.gartner.com/newsroom/id/2209615.

[5] Murazzo, Rodríguez, Villafañe, González. “Análisis de grandes volúmenes de datos en el Cloud”. I Jornadas de Cloud Computing. La Plata, 17 al 19 de junio de 2013.

[6] Srinivasa Rao, Nageswara Rao, Kusuma Kumari, “Cloud Computing: An Overview”. Queue 7, 5 (Jun. 2009), 3-4.

[7] Murazzo, Rodríguez, Segura, Villafañe. “Desarrollo de aplicaciones para Cloud Computing”. CACIC 2010. Morón. Oct. 2010.

[8] Google, Inc. **“**Google App Engine (GAE)”. http://code.google.com/intl/es-ES/appengine.

[9] Amazon.com, Inc. “Amazon EC2”. http://aws.amazon.com/ec2.

[10] Microsoft Corporation. “Windows Azure”. http://www.microsoft.com/windowsazure.

“ASLX”: Semantic Analyzer for programming languages based on XML

Gabriela Yanel.Buffa, Franco Gaston.Pellegrini
Licenciatura en Ciencias de la Computación
Universidad Nacional de Río Cuarto

Abstract. XML (Extensible Markup Language) is a language used to structure information in a document or any file containing text. An XML document is "well formed" if respects their *basic syntactic structure*, that is composed of elements, attributes, and specified as XML comments. Some ways to verify that an XML file is either formed is by: *Document Type Definition (DTD)* and *XML Schemas*. But these types of verification not include relevant aspects, such as, the semantic relationship between content attributes and tags, declarations of fields and check the scope of variables, among others. This raises the need for a tool for verifying programming languages based on XML in the semantic aspect. In addition to analyzing the scope and usability of the tool was used as a case study the NCL declarative language.

Acknowledgements

Mainly, we want to thank our families, who gave us the opportunity to study, to be where we are, who supported us from day one, whether in our achievements, failures and mainly because without their help it would have been impossible to reach these instances.
Also, thank you to those who helped us during this hard work, our Director Mg. Marcelo Arroyo, for proposing to address the issue, and our Director Dr. Francisco Bavera, being both present in each stage, mainly in every mistake, and without whom this work could not have been carrying out.
For my part (Yanel) I want to thank the angel that I have in heaven, "My Grandmother" or rather my second Mom, which helped me at the time to be here today and then from somewhere nursed and gave me strength each day to fulfill this achievement.

1 Introduction

This work aims to develop a solution to validate semantically languages programming or the like based on XML.

In **XML** (Extensible Markup Language), XML files are text files where symbols "greater than" and "less than" are used to delimit the brands that give structure to document. Each brand has a name, let's see an example:

The brand <figure>, may have one or more attributes:

```
file="foto1.jpg" <figure tipo="jpeg">
```

In this case you have two attributes, "file" and "type".

The attributes have values that have to be in quotes or apostrophes, the marks cannot be left open, i.e., for each brand, for each brand, <*figure*> for example, there should be a mark </ *figure*> indicating where it ends for the content of the brand. That a document is "well formed" only refers to its *basic syntactic structure* is say, which is composed of elements, attributes and XML comments as specified to be written. Now, each XML application, i.e., each language defined with this technology, need to specify what exactly the relationship which must hold between items present in the document.
Some ways to verify that an XML file is well formed, is by:

1. *Document Type Definition or DTD:* Defines the types of elements, attributes and entities permitted, and can express some limitations to combine. XML documents that conform to your DTD are called valid.
2. *XML Schemas:* A Schema is similar to DTD defines which may contain elements XML document, how they are organized, what attributes and what type may have their elements.

Between the two types of verification, there are several fields that are not covered, and they are very useful for verify programming languages based on XML:

1. Semantic relationship between content attributes and tags.
2. Importing variables (and type) of another file.
3. Semantic verification of file formats to variables.
4. Statements Scopes and variable scope verification.
5. Conditional checks the contents of an attribute.

Some of these can be done in a very basic way using regular expressions in XML Schemas, but the problem is limited to a particular attribute and *not* the relationship between several.

This raises, the need for a tool for verifying programming languages based in XML or similar in the semantic aspect. In short, to develop an application that can extend verification of validity that provides XML with the above points and more.

To apply our tool to a real problem, it was decided to use it to validate semantically source code of declarative language NCL.[1]. Create NCL applications is not very difficult for people in the field of programming, requires prior learning which is difficult because such language has only one basic data type ("String") which is used to reference to any basic entity or media item (video, image, etc.), and interpretation tools Ginga-NCL do not provide basic facilities of a modern compiler for relevant reports to the user from their errors.

[1] Gomes Soares, Luis Fernando: "Programando em NCL": Desenvolvimento de aplicações para middleware Ginga, TV digital e Web. (2009).

In this way and as previously mentioned, according to our knowledge, "ASLX" is the first semantics tool for the NCL language and one of the few semantic tools for XML-based languages or similar.

1.1 Development Objectives:

Considering that:

- It is very complex to find errors in a program if the compiler does not help.
- Find and correct errors confusing error outputs watching is very productive.
- That does not even cover all aspects of verification to verify if a file XML is well formed.

Main objectives are proposed, creating a semantic validator detected in the lower long as possible the different semantic errors and / or syntactic language can have a programming based on XML.

The tool can display a detailed report of the errors, thus this tool can be of great help to many developers at the time to correct their applications, since providing a detailed analysis may go directly to the point of failure and fix it. Another feature that should have the program, is to be used as a library (so that for example an IDE can leverage the speed and advantages of this solution).

Moreover, considering that Ginga-NCL does not verify semantics, such language was used as a case study and test, to thereby verify the semantic concepts (like well as in the syntax) that verifies Ginga-NCL alone.

1.2 Using a Scripting Language

A scripting language is a type of programming language that is generally interpreted (as opposed to compile). A script can be seen as a program that can accompany a document HTML or contained inside. The scripts remain in its original form (your source code Text) and are interpreted command by command whenever running.

Scripting language features:

1. Scripts are usually written more easily, but at a cost of its execution.
2. Usually implemented with interpreters rather than compilers.
3. They have strong communication with components written in other languages.
4. The scripts are usually stored as plain text.
5. The codes are usually smaller than the equivalent in the language of compiled programming.

In developing this tool, the Scripting language to create semantic rules are designed with the aim of providing comprehensive solutions to different problems that only verify semantics of a single language (eg Ginga-NCL). Through this mechanism, it can be adapted our program to verify semantics of programming languages based on XML.[2]

A Script begins with a tag called *"semanticParser"* which has reference to the schema script. Must also be specified using the attribute "fileFormat" a list (separated by commas) of extensions this script supports to validate.

Then, within the tag *"rules"* specifies all the rules. For this example we want to set a set of rules that should be applied only to tags labeled with the name "media". With this objective stated in the "name" attribute on the tag "tag" value "medium".

[2] For more information on scripting languages and their use in this tool, refer to Chapter II of the full report of the developed tool "ASLX".

2 Design

2.1. General Design

The development process is centered on a defined architectural with rules developed in a script based in XML. First developed an XML schema *"ScriptParserSchema.xsd"*, in the defined a language for creating scripts with the semantic rules to check.

Given the complexity, it was decided to design incrementally. Our program takes the script and a code based on XML as a parameter, follow the rules defined in the script and try to validate the XML code reporting errors or warnings.

Subsequently, we studied the specific failures of Ginga-NCL for rules and / or tools needed to incorporate.[3] In the case of Ginga-NCL designed a script, *"semanticCheck.xml"* with all the rules semantics of the same so that by giving one or more files with NCL code, the program can verify and create a report. This is defined by regular expressions and rules (applied to "Tags") the conditions to be met by any program NCL to be a valid application free the semantic errors.

The program was divided two sections: "*Parser"* and "*Validator"*. If no syntax errors, the Parser interprets the rules defined in the script and generates a Validator results. This Validator is used to check the rules on one or more files with the format specified in the script.

The basic set of rules that recognizes Parser is:

1. Existence of an attribute with a given name.
2. Right contents of an attribute or content right relationship between various attributes.
3. Check existence of local or external files.
4. Creating objects / variables referenceable.
5. Type checking (for references to objects / variables created).
6. Include code from other files.
7. Check text content of an XML node.
8. Conditional validation rules.

2.2 Structure and Course of an XML file

The design for parsing XML files based on the use of API *"SAX"* and *"DOM"*. These APIs are two tools used to analyze XML and define the structure of a document, although there are many others.

2.2.1 Using DOM and SAX in the design

Because SAX is more complete for information storage, and DOM is simpler to use since there is no need to implement a tree like structure data, we decided to combine both.

By SAX modify the XML file parsed nodes and we add location (for which line of code is the node). Using DOM created the tree data structure to tour files easily without having to implement anything. Also used SAX to validate XML files as one XML Schema, useful to verify that the script and files to be verified with a base stable.

[3] Gomes Soares, Luis Fernando. "Programando em NCL": Desenvolvimento de aplicações para middleware Ginga, TV digital e Web. 2009.
Associacao Brasileira de Normas Técnicas. "ABNT NBR 15606-2": Televisão digitalterrestre – Codificação de dados e especificações de transmissão para radiodifusão digital.

3 Implementation

3.1 Introduction

As programming language Java 7 was chosen only by two important features:

1. Write software on one platform and run it on virtually any other platform.
2. Simple to produce code quantity and quality (productivity).
3. *SAX* and *DOM* are included in the standard library.

For technical implementation details, refer to the technical documentation of the code (Javadoc or code). In this section we discuss only general details of the source code.

3.2 Implementation Details

The entire project was created under the IDE NetBeans[4], which facilitates file *"build.xml"* to compile your code without the need to install this IDE.

Using Apache ant[5] can run the file "build.xml" and automatically generate Javadoc documentation and executable JAR extension.

Since the design is very simple because it uses DOM and SAX to parse the scripts and validate files, only discussed in this section general details of the most important classes of program. For details, refer to the technical report or the source code.

3.2.1 Parser (ScriptParser.java)

You cover all valid nodes recursively with the defined rules (script) using an algorithm.[6] Whenever the node name matches the name of a reserved word of the script, it acts according to the meaning of it by creating rules, ER-Ex (patterns) or containers of rules associated with a tag (TagValidator.java). To facilitate the implementation and script writing, all white characters of ER-Ex will be ignored. If needed them, using suitable escapes characters.

To solve the macros of the ER-Ex, the parser tries to do the translation of the macro only one time, and stores the results. Since a macro always comes down to the same ER, is efficient only calculate this reduction only once. This allows more efficient management of memory and speed parsing. Also added is a limit to the calculation of the macros, as solving recursively ill-defined macro can stop the program in a state of infinite calculation. After obtaining the ER and rules, this returns a Validator

3.2.2 Container Rules (TagValidator.java) and Validator (Validator.java)

Represents a set of rules only applicable to a particular tag. They are stored rules created by the script ready, for easy retrieval. It all nodes recursively valid XML file being checked by the algorithm, and if the name of some of the nodes are in the container list of rules (TagValidator) apply all the rules contained therein to validate that node / tag.nodes in the container list of rules (TagValidator) apply all the rules that thiscontains to validate that node / tag.

[4] http://netbeans.org
[5] http://ant.apache.org
[6] Algorithm described in the section "General Program Design", Chapter II of the full report of the developed tool "ASLX".

4. Testing

4.1 Introduction

The tests performed to verify the correct operation of the program, were of “Black Box”. *JUnit* was used to design a set of test cases and found as it implemented would program the outputs are exactly as expected. The most complete test case is the demonstration of the script with validation rules semantics for Ginga-NCL language which uses most of the functionality of the program in a real case. All tests are described next to the source code and technical reports.

4.2 Optimizations

The first tests performed were of optimization. The evaluation was specifically related to the response time of the main data structures and management of ER-Ex. We implemented five versions of the program, which were evaluated with 4 different scenarios.

The 5 versions had the following characteristics:

Version	*Characteristics*
1	No optimizations
2	The ER equal solved only 1 time
3	The ER-EX Macros are solved only 1 time
4	Macros and ER are stored in HashMap instead of SortedMap (TreeMap)
5	HashMap across data structure that requires searchesin Validator and Parser

Each version has included optimizations of the previous version.

All performance measures were obtained on a Notebook with an Intel ® Core ™ i7-2630QM CPU@2.00GHz × 8 and 4Gb DDR3 RAM under Ubuntu 10.04 64bit (Linux kernel 3.2.0-25-generic).

The results were:

Version 1

Stage	The average time (ms)	ER created	ER reused	Macros analyzed	Macros reused
1	13	4	0	7	0
2	938	5000	0	0	0
3	2477	5000	0	0	0
4	17202	20	0	2334634	0

Three versions were made with the same number of stages, which will be omitted in this report.[7]

[7] For more details refer to Chapter V of the full report of the developed tool "ASLX".

The following shows in detail the latest version and optimizations with respect to the first:

Version 5

Stage	The average time (ms)	ER created	ER reused	Macros analyzed	Macros reused
1	6	4	0	4	3
2	700	5000	0	0	0
3	166	100	4900	3	5049
4	2717	10	10	104	2

This version besides being faster (in some cases) use less RAM than other versions, reason why it was chosen as the implementation.

4.3 JUnit and Test Scenarios of Script Language

For proper implementation incrementally, we chose to go to implement the program validating it through the JUnit[8] framework. JUnit return to class method successfully passed the test, if the expected value is different from that returned during the execution method, JUnit failure to return a corresponding method. JUnit is also a means of controlling regression testing, necessary when part of the code has been modified and you want to see the new code meets the above requirements and has not altered its functionality after the new amendment.

4.4 "Testing Case

4.4.1 Syntax Errors in XML Schema of Ginga-NCL.

For the testing phase, extracted oficial book[9], a total of 33 Ginga XML schemas-NCL. A Below is a detail of the scheme that had errors, the page which can be found in oficial book, a detail of the errors and possible solutions that could be implemented, taking into tool has developed "ASLX" to thus be valid applications.

"CausalConnector.xsd": The scheme is located on page number 48 of the official book.

Code where the error (line 40):

```
<complexType name="nclType">
     <complexContent>
          <restriction base="structure:nclPrototype">
               <sequence>
                   <element ref="structure:head" minOccurs="0"
maxOccurs="1"/>
                   <element ref="structure:body" minOccurs="0"
maxOccurs="0"/>
               </sequence>
          </restriction>
     </complexContent>
</complexType>
```

[8] http://www.junit.org//

[9] Gomes Soares, Luis Fernando. “Programando em NCL”: Desenvolvimento de aplicações para middleware Ginga, TV digital e Web. 2009.

Error type:

```
cos-particle-restrict.2: Forbidden particle restriction:
'choice:all,sequence,elt'.
```

Solution
Defining the maximum of occurrences (maxOccurs) body structure is 1 instead of 0.

4.4.4 Syntactic errors in examples official Ginga-NCL.

We extracted all the code examples from the book mentioned above for official to use them as evidence. In this way we obtained a total of 20 examples which first syntactic errors were corrected.

Beyond that the tool developed to detect syntax errors to achieve fully valid applications, as the main objective is the semantic, are omitted in this report[10].

4.4.5 Semantic errors in examples official Ginga-NCL.

Were detected errors in the following examples Ginga-NCL. These errors were found by the validator semantic created. As noted, various examples were no syntax errors, if they have them in the semantic aspect. In the absence of the tool created, these examples would be considered valid when in fact they are not.

"Example 1": This example is the initial release of the NCL document "O Primeiro João". This example can be seen on page number 51 of the official book.

Code where the error (line 36):

```
file:Ejemplo1.ncl: line:36: The symbol "conEx#onBeginStartDelay" is
not
declared in this scope.
file:Ejemplo1.ncl: line:45: The symbol "conEx#onBeginStart" is not
declared in this scope.
file:Ejemplo1.ncl: line:49: The symbol "conEx#onBeginStart" is not
declared in this scope.
file:Ejemplo1.ncl: line:53: The symbol "conEx#onEndStop" is not
declared
in this scope.
```

Solution
CausalConnBase.ncl Define a new file, which contains each of the connectors for compiling the examples present in Ginga-NCL. In this case specifically defined in the file, the connectors *"onBeginStartDelay"*, *"onBeginStart"* and *"onEndStop"*.

[10] For more details refer to Chapter V of the full report of the developed tool "ASLX".

Below is a table show examples present in the official book of GINGA_NCL possessing semantic errors:

Example	Description	Error line	Error type	Solution
Example 3.15 (Page 56)	Application to sync by interaction	41-50-54-58-63-67-77	Elements are not declared In this scope.	Define the file causalConnBase.ncl the connectors.
Example 3.19 (Page 63)	Application to Reuse	43-47-57-64-73-77-81	Elements are not declared In this scope.	Define the file causalConnBase.ncl the connectors.
Example 3.22 (Page 67)	Application to channel interactive	47-51-62-69- 78	Elements are not declared in this scope.	Define the file causalConnBase.ncl the connectors.
Example 3.27 (Page 73)	Application with content adaptation	60-64-75-82-91- 95-99	Elements are not declared in this scope.	Define the file causalConnBase.ncl the connectors.
Example 3.32 (Page 79)	Application-controlled interactive ads.	53-60-66-70-98-100-103-114-121-130- 134-138	Elements are not declared in this scope. Invalid value in attribute "role".	Define the file causalConnBase.ncl the connectors. Give a valid value to the attribute "interface" and to attribute "role".
Example 3.37 (Page 85)	Application purposes animations and transitions	60-67-77-80-85 90-120- 127-136-140	Elements are not declared in this scope. Invalid value in attribute "interface".	Define the file causalConnBase.ncl the connectors. Give a valid value to the attribute "interface"
Example 3.45 (Page 95)	Application menu with navigation keys	72-79-89-117-122-133-139-167-174-185-194-198-207-213	Elements are not declared in this scope.	Define the file causalConnBase.ncl the connectors.
Example 3.52 (Page 105)	Application in order NCLua	60-74-81-91-120-122-125-136-142-156-175-185-199-208-212-221-229	Invalid value in attribute “end” y “role”. Elements are not declared in this scope.	Give a valid value to the attribute "interface" and attribute “role” Define the file causalConnBase.ncl the connectors
Example 7.1 (Page 144)	Example showing the interactivity of a "button" for a fixed time	18	The symbol "dTVtelaInteira" is not declared in this scope <even forward>.	Writing the shape descriptor correct.

Developed tool also found errors in the examples: "10.12" (page 199), "10.14" (page 202), "12.4" (page 234), "14.4" (page 259), "14.6" (page 268), "15.3" (page 272) "15.6" (page 274).
In case you want to see in detail this type of error, refer to Chapter V of the full report tool developed "ASLX".

References

1. Gomes Soares, Luis Fernando: “Programando em NCL”: Desenvolvimento de aplicações para middleware Ginga, TV digital e Web, (2009).
2. Associacao Brasileira de Normas Técnicas: “ABNT NBR 15606-2”: Televisão digitalterrestre – Codificação de dados e especificações de transmissão para radiodifusão digital.
   Parte 2: Ginga-NCL para receptores fixos e móveis – Linguagem de aplicação XML para codificação de aplicações (2007).
3. Gomes Soares, Luis Fernando and Ferreira Rodrigues, Rogerio: “Nested Context Language 3.0 , NCL Digital TV Profiles” (2006).
4. Soares, Luiz Fernando Fernando Gomes: “Construindo Promgramas Audiovisuais Interativos Utilzando a NCL 3.0 e a la Ferramentea Composer” (2007).
5. Zambrano, Arturo: “Introducción a la TV Digital Interactiva y Ginga.ar”- Universidad Nacional de La Plata. La Plata – Argentina.
6. Balaguer Federico & Isasmendi, Leonardo: “Desarrollo de Aplicaciones para Televisión Digital”. Universidad Nacional de Río Cuarto. Rio Cuarto - Argentina (2011).

Sistema Guía para Personas con Deficiencia Visual

Pablo Richard, Daniel Richard, Marcos Aranda

Universidad Nacional de Catamarca
{pablorichard885, danielrichard86, marcos_dario_1}@hotmail.com

Resumen. En este artículo se presenta un sistema guía para personas con deficiencia visual basado en la construcción de un dispositivo electrónico de sensado, una aplicación móvil cliente y la comunicación y sincronización entre las partes. Este producto ayudará a las personas que padecen esta patología a desenvolverse independientemente en su entorno, ya que detecta la presencia de objetos que obstaculizan el desplazamiento. La aplicación corre en un dispositivo móvil y se adapta a las necesidades de cada usuario. El sistema incluye la construcción de un dispositivo con sensores ultrasónicos de distancia desarrollado con tecnología PIC, el cual se comunica con una aplicación móvil desarrollada en Java para Android que se ejecuta en un teléfono celular. Se utiliza el paradigma de programación orientado a objetos para el desarrollo de la aplicación debido a que permite la reutilización y extensión del código. De acuerdo a las pruebas funcionales realizadas, es factible la implementación exitosa del sistema.

Palabras Claves. Aplicación móvil, sistema guía para deficiencia visual, PIC, sensores ultrasónicos de distancia, distanciómetro, orientación a objetos.

1. Introducción

Los problemas de visión acarrean a las personas que los padecen una serie de obstáculos que, para ellos, se transforman en grandes retos a superar. Esto se pronuncia aún más en una sociedad muy dependiente de los estímulos visuales en la vida cotidiana (letreros, televisión, imagen social, computadoras, etc.). Es necesario diferenciar a las personas que sufren estos padecimientos en algún momento de su vida en forma aguda, de aquellas que lo padecen de nacimiento o de muy temprana edad; para los primeros es mucho más difícil porque tienen que adaptarse a un nuevo estilo de vida.

Los avances tecnológicos y la evolución de la computación brindan recursos para poder desarrollar herramientas que ayuden a estas personas con algunos de sus problemas cotidianos.

El desarrollo de este proyecto tiene como objetivo brindar un sistema que le ayude a las personas con deficiencia visual a desenvolverse de una mejor manera en su entorno. Para ello se realizó la construcción de un dispositivo electrónico con microcontroladores y sensores, que es capaz de capturar y manejar información del medio (hardware del sistema), que se relaciona con la aplicación (software del

sistema) para que la misma haga uso de esta información, lo cual dio el soporte necesario para poder obtener los datos suficientes del entorno donde se mueven.

La conexión entre el dispositivo electrónico y la aplicación se realizó mediante Bluetooth lo que permite la comunicación y sincronización de las partes del sistema, para ello se implementaron librerías para manejo de la tecnología anteriormente mencionada y se estableció un protocolo de comunicación para su uso.

La aplicación reconoce el medio o el entorno fijo y también es transportable, permitiendo el desplazamiento del usuario.

Actualmente, los dispositivos móviles (celulares, tablets, laptops, etc.) son capaces de correr aplicaciones que son de uso cotidiano en la vida de cualquier persona. Esto es lo que motivó a desarrollar software ejecutable sobre estos dispositivos.

El sistema detecta la ubicación de los objetos alrededor de personas con deficiencia visual y los guía en decisiones referidas a su movilidad. Está compuesto por: una aplicación móvil y un dispositivo de sensado ultrasónico de distancia. La aplicación corre en un dispositivo móvil (celular, notebook, etc.) que tenga sistemas operativo Android con versión 2.5 en adelante, e indispensablemente una conexión Bluetooth, ya que se conecta al dispositivo de sensado a través de Bluetooth. El dispositivo se desarrolló utilizando microcontroladores PIC.

2. Desarrollo del sistema

El sistema abarca la construcción del dispositivo electrónico, del desarrollo de la aplicación móvil y la comunicación entre el dispositivo y la aplicación.

Para el desarrollo de este sistema se emplearon diversas tecnologías para las diferentes partes que componen el sistema.

El dispositivo electrónico está compuesto por un microcontrolador PIC 18F4550 que es la unidad central del mismo, un sensor ultrasónico Parallax, un display LCD y un módulo bluetooth para establecer la comunicación con la aplicación del dispositivo móvil. Para la implementación de la aplicación, se utilizó un dispositivo móvil LG L7, pero puede utlizarse cualquier dispositivo móvil que posea Android desde la versión 2.2 en adelante y una comunicación bluetooth.

A continuación se describe la construcción del dispositivo electrónico y de la aplicación.

2.1. Diseño y construcción del dispositivo electrónico

En este apartado se describirán toda la ingeniería que se aplicó para diseñar y construir el dispositivo electrónico *distanciómetro* encargado de medir la distancia entre el usuario y el objeto.

Diseño general del sistema de medición de distancias por ultrasonido:
En la figura 1 se presenta un diagrama en bloques, el cual describe en forma general los elementos principales del distanciómetro.

- Oscilador: El cuál es el encargado de entregar al sistema la onda cuadrada, cuya frecuencia es ultrasónica de 40 KHz.
- Modulador: Tiene la función de generar ráfagas de ondas ultrasónicas, o sea, funciona como una llave electrónica que deja pasar una cantidad de pulsos limitados provenientes del oscilador.
- Amplificador: Ésta etapa adapta las ráfagas ultrasónicas a niveles de tensión adecuados para el trabajo del transductor transmisor. En el caso del receptor, se amplifica la señal para que pueda ser procesada adecuadamente por el sistema de control, dado que la misma llega disminuida al receptor.
- Sistema de control: es el cerebro del sistema, sobre el cual corre el programa que controla todos los bloques.

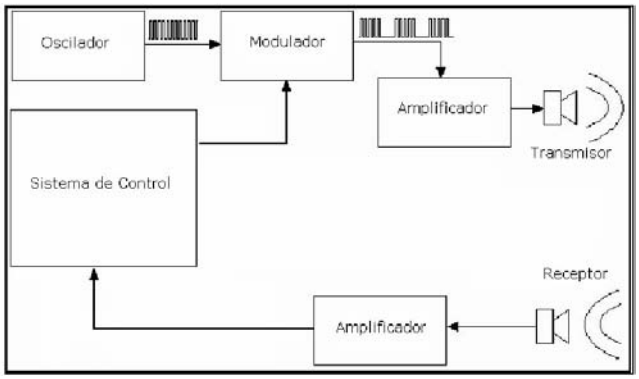


Figura 1. Diagrama de bloques de un distanciómetro ultrasónico.

Circuitos para transmisión y recepción de ultrasonidos para la medición de distancias:
En el circuito transmisor, se generan ráfagas de 40 KHz con duración de 200 microsegundos, cada 65 microsegundos, al detectar la onda reflejada se genera una interrupción la cual detiene un temporizador (*timer*) de 16 bits del microcontrolador.

Transmisor ultrasónico:
El transductor utilizado para la transmisión es el piezoeléctrico N1076, controlado por el microcontrolador PIC12C508, el cual se encarga de enviar el tren de pulsos de 40KHz para que el cristal emita la frecuencia de ultrasonido deseada. Dado que el transmisor empleado soporta una tensión de entrada de hasta 20 Vrms, se incluye el acoplamiento con el componente ST232 entre el microcontrolador y transductor el cual permitirá una tensión de entrada al emisor de aproximadamente 16 V [1].

Como se mencionó, el microcontrolador es el "cerebro" de la operación. En él corre el programa que permite la generación de las ráfagas ultrasónicas.

En la figura 2 se muestra el esquema del transmisor, representado mediante la utilización del software de diseño y simulación electrónica MULTISIM:

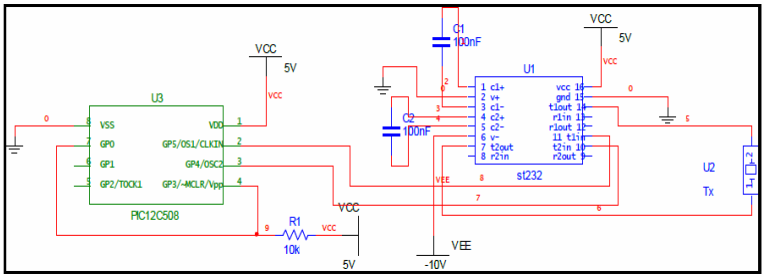


Figura 2. Diagrama de conexión del circuito emisor ultrasónico. Fuente: www.uhu.es/antonio_peregrin/iaic_abierto/SRF04.PDF.

Receptor ultrasónico

El receptor se compone de dos circuitos amplificadores de señal y un circuito de detección.

La señal es recibida por el sensor receptor y amplificada 576 veces en dos pasos por 2 amplificadores por 24. En esta etapa se hace uso de los circuitos integrados LM1458 cuyo ancho de banda es 1 MHz, cuya máxima ganancia corresponde a la frecuencia de 40 KHz.

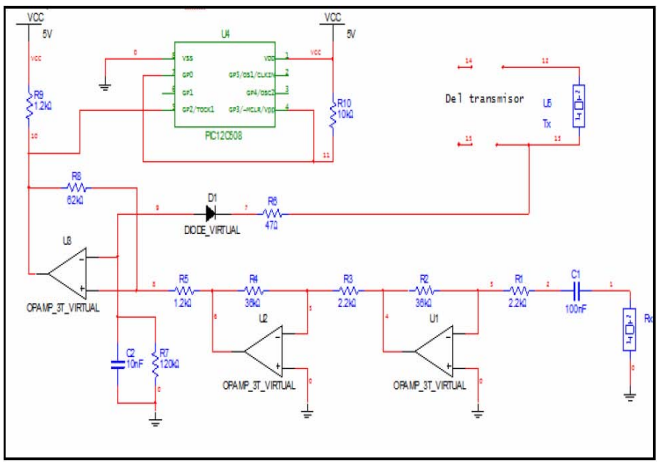


Figura 3. Diagrama de conexión del circuito receptor ultrasónico. Fuente: www.uhu.es/antonio_peregrin/iaic_abierto/SRF04.PDF.

Como puede verse en la figura 3, en la salida de las etapas amplificadoras hay un comparador (LM311), el cual tiene como entradas la señal proveniente del receptor (entrada positiva) y la señal del transmisor (señal negativa). Se agrega además una realimentación positiva por medio de la resistencia de 62 KΩ para agregar histéresis, dando estabilidad a la salida del comparador.

Por último, el resultado de la comparación es procesado por el microcontrolador PIC 12C508, el cuál detiene el conteo del temporizador cuando ésta señal llega.

Funcionamiento

Tal y como se muestra en el diagrama de tiempos de la figura 45, el funcionamiento es descrito, por parte del usuario y por medio del microcontrolador PIC 18F4550, un pulso de disparo o "*trigger*" de 2 microsegundos que inicia la secuencia. Por medio del transductor Tx se transmite un tren de pulsos o "*burst"* (ráfaga) de 200 microsegundos a 40KHz. En ese momento, el microcontrolador PIC18F4550 debe cambiar su condición de salida a entrada para esperar por el mismo pin la señal de "eco", por lo tanto en este momento se envía al PIC 18F4550 un "1" lógico. Cuando la cápsula receptora recibe la señal transmitida como consecuencia de haber rebotado en un objeto (eco), se envía al PIC18F4550 de nuevo un "0" lógico. En este microcontrolador se realiza la medición de la duración del pulso de ésta señal, es decir, el tiempo en que la señal recibida anteriormente se mantiene a "1" [2].

Con objeto de que el sensor se estabilice, se deja un pequeño intervalo de tiempo de 10ms como mínimo entre el momento en que la señal de eco pasa a "0" y un nuevo pulso de disparo que inicie el siguiente ciclo o nueva medida.

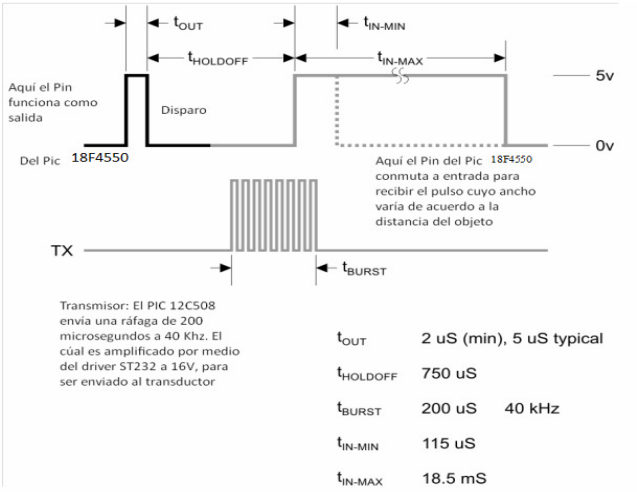


Figura 4. Diagrama de tiempos del funcionamiento del sensor ultrasónico. Fuente: www.parallax.com/dl/docs/prod/acc/28015-PING-v1.3.pdf

Como muestra el gráfico, existen 2 límites de tiempo: el tiempo mínimo de 115 microsegundos, es el fijado por el software y corresponde a la medida de 3 cm aproximadamente, puesto que medidas por debajo de los 3 cm provocan una serie de errores derivados del acoplamiento entre las propias cápsulas emisor-receptor del módulo, por lo que es muy difícil distinguir si la señal recibida es consecuencia de dicho acoplamiento o del eco recibido [3].

Por otra parte el tiempo máximo es de 18.5 milisegundos, el cual corresponde a una distancia aproximada de 3 m. Este límite también impuesto por el programa + 10 milisegundos de resguardo, es el tiempo que se debe esperar para que el PIC18F4550 conmute nuevamente su pin a entrada para generar un nuevo disparo de activación de secuencia.

Para mayores distancias nos podemos encontrar con problemas derivados de la dispersión del haz ultrasónico o de múltiples rebotes que pudieran generarse.

Funcionamiento PIC 18F4550

El programa que corre en el PIC, comienza con la generación y envió del pulso de disparo de inicio de secuencia de 10 microsegundos de ancho a través del pin 20, para que se genere el *burst*. Inmediatamente el pin 20 conmuta su configuración convirtiéndose en entrada y se prepara para recibir el pulso de eco, cuyo ancho varía dependiendo de la distancia en la que se encuentre el objeto. Durante ese tiempo se activa un temporizador interno (*timer*), que lleva el conteo de la duración del pulso de eco. A continuación esta cuenta es almacenada por el PIC, donde se produce la multiplicación por factores de escala con el fin de llevar esa medición a centímetros. Luego este dato es mostrado en el display LCD.

Por último el programa retorna a la rutina de generación del disparo para iniciar una nueva secuencia de medición produciendo un bucle constante.

A continuación se muestra una simulación del dispositivo electrónico en el software ISIS Proteus que nos permite no solo plasmar nuestro diseño, si no cargar el programa en el PIC y simular su funcionamiento.

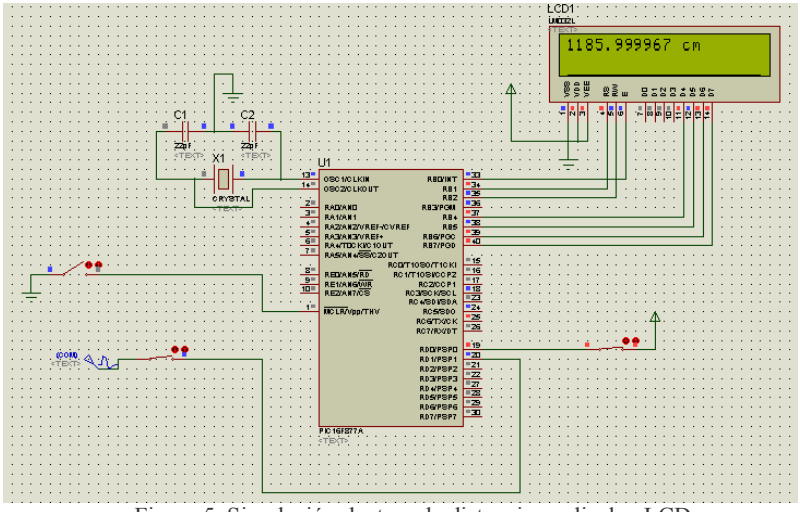


Figura 5. Simulación, lectura de distancia en display LCD.

Montaje en placa experimental

Como se observa en la figura 5, la simulación de la funcionalidad principal del dispositivo electrónico, calcular la distancia entre el usuario y el objeto, trabaja de forma correcta entonces plasmamos nuestro diseño en una placa de prueba obteniendo un resultado físico del mismo. Esto se puede apreciar en la figura 6.

Figura 6. Imagen del prototipo.

Construcción del dispositivo segmento para la comunicación

Lograda la función principal del distanciómetro, adaptaremos el módulo Bluetooth RN-42 al prototipo para establecer la comunicación inalámbrica con la aplicación. Este es un módulo de clase 2 que permite la comunicación inalámbrica por enlace bluetooth con un alcance de hasta 20 mts posee un protocolo de saltos de frecuencia que le permite actuar en ambientes con interferencias, un protocolo de control de errores para la comunicación por bits de paridad y para la transferencia de datos utiliza la conmutación de paquetes y circuitos.

En la figura 7 se puede ver el dispositivo electrónico terminado montado en una placa fija.

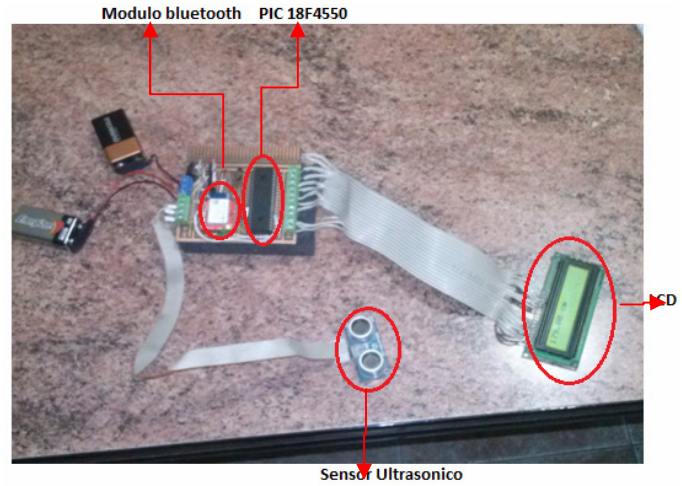


Figura 7. Dispositivo en placa fija.

2.2. Comunicación entre el dispositivo y la aplicación

En este apartado se describirán las librerías utilizadas, tanto a nivel software como firmware para realizar la comunicación entre el dispositivo electrónico denominado distanciómetro y la aplicación.

Por parte del dispositivo electrónico se utilizaron librerías para la comunicación serial asincrónica ya que el modulo bluetooth RN-42 se conecta por los pines del PIC de comunicación serial y actúa como un adaptador para la transmisión por aire.

Por parte de la aplicación se implementaron las librería que proporciona el api de google para comunicación bluetooth con android que disponen un conjunto de clases numerosas que permite manejar los recursos para llevar a cabo un enlace bluetooth con el dispositivo móvil. [5,15].

2.3. Desarrollo de la aplicación

La aplicación trabaja con la información recibida del Distanciometro; por lo tanto cuando éste inicia se debe conectar con el mismo. El distanciómetro a través del modulo bluetooth RN 42 le envía el dato de la distancia que se encuentra el mismo ante un objeto, el programa que corre en el dispositivo móvil recibe el dato por en el enlace bluetooth que establece con el distanciometro y analiza este valor, emitiendo un sonido o una vibración, indicándole al usuario en la situación que se encuentra para que decida como actuara respecto a la misma.

Para esto se contemplaron 3 posibles casos:

- Sonido 1 o vibración 1: esta situación se dispara cuando el valor de la distancia recibida entre el distanciómetro y el objeto no representa peligro alguno para el mismo.
- Sonido 2 o vibración 2: esta situación se dispara cuando el valor de la distancia mencionada representa un cierto grado de peligro para el mismo.
- Sonido 3 o vibración 3: esta situación se dispara cuando el valor de la distancia mencionada representa peligro para el mismo.

El programa realiza esta acción cada un segundo mientras el usuario no se desconecte del dispositivo o cierre la aplicación, esto se lleva a cabo mediante un timer.

La aplicación permite la configuración de los parámetros de distancia que se ajustarán a los casos de decisión posible adaptándose a las necesidades de cada usuario a la hora de transportarse.

En cuanto al desarrollo de la aplicación, se siguió un método orientado a objetos. En la figura 8 se presenta un diagrama de las clases relevantes que intervienen en el sistema, que definen el contexto del mismo. En la figura 9 se presenta un diagrama de secuencia que demuestra la forma que interactúan los procesos y los actores que intervienen en el sistema.

En la figura 10 se muestran la interfaces del sistema, que consta de 3 pantallas sencillas e intuitivas.

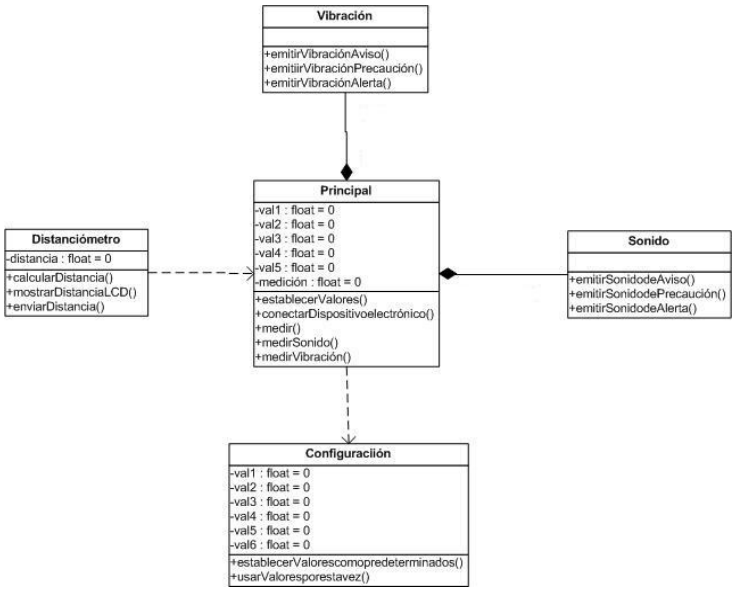


Figura 8. Diagrama de clases del sistema.

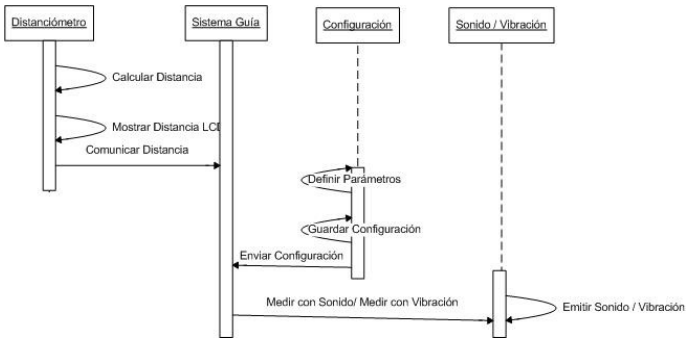


Figura 9. Diagrama de interacción.

La primera pantalla de la figura 10, es la de inicio del sistema que está dispuesta por un menú, "Comenzar" nos envía al sistema de guía con valores predeterminados, "Configuración" nos permite configurar los parámetros de distancia y alertas con el que correrá el sistema de guía y "Salir" permitirá cerrar la aplicación y liberar recursos del dispositivo móvil. La Segunda pantalla es la de Configuración del sistema acá se puede configurar los valores de distancia que se adaptaran de acuerdo a la necesidades del usuario es decir a sus patrones de movimiento. La tercera pantalla es la Principal dispone un sistema de guía intuitivo, compara los valores que son enviados por el distanciometro con los que configuro al usuario y emite un sonido o vibración que indica al usuario la situación en que se encuentra.

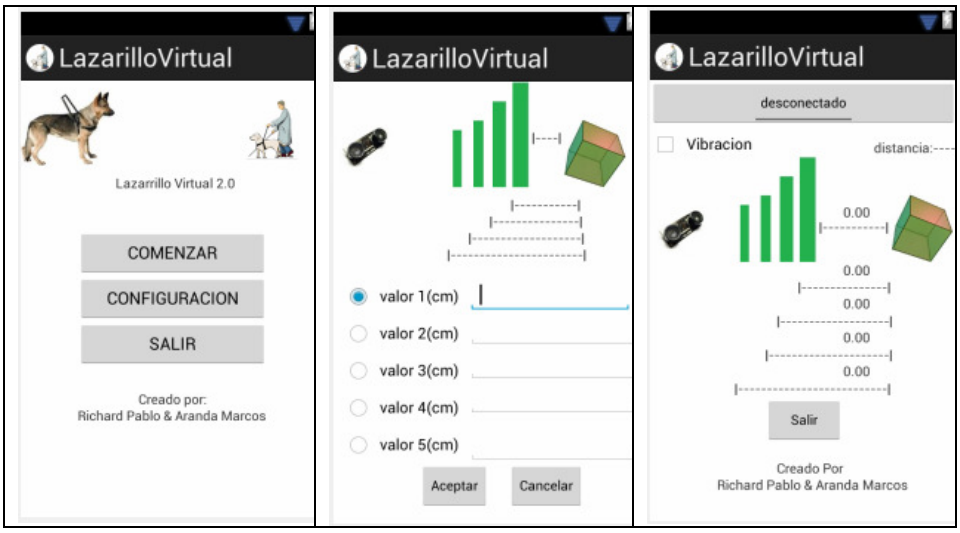


Figura 10. Pantallas de inicio, de configuración y de uso del sistema.

3. Conclusiones

Se concluye que, con la utilización de nuevas tecnologías y la evolución de la computación, se obtuvieron los recursos necesarios para desarrollar una aplicación para dispositivos móviles con sistema operativo Android y construir un dispositivo electrónico seguro y robusto, que actuando en forma conjunta conforman un sistema de guía que resuelve la problemática de poder transportarse de forma independiente y segura que padecen las personas con deficiencia visual.

El diseño del sistema no solo cumple con los objetivos del proyecto si no también brinda una solución simple, ya que los dispositivos móviles son un elemento común en la vida cotidiana de cualquier persona, además este diseño le permite al sistema crecer potencialmente teniendo en cuenta todas las funcionalidades que proporciona un dispositivo móvil. De esta manera, se podrá satisfacer otras demandas o necesidades del usuario a futuro.

4. Referencias

1. Angulo J., Microcontroladores PIC: Diseño Práctico de Aplicaciones. 3ª Ed. Madrid -España: McGray-Hill, (2003).
2. Corrales S., Electronica Práctica con Microcontroladores PIC. Quito-Ecuador: Imprenta Grafica, (2006).
3. Daniel Benchimol, Microcontroladores. Buenos Aires: Fox Andina, (2011).
4. Donn Felker y Joshua Dobbs (2011), Android application development for dummies. Indiana: Wiley Publishing, Inc.

Automatización en la Captura de Datos para el Modelado de Flujo Vehicular

Julio Monetti, Mariana Brachetta y Oscar León
Universidad Tecnológica Nacional – Facultad Regional Mendoza
Rodriguez 273- Mendoza - Argentina
{jmonetti,mbrachetta,oleon}frm.utn.edu.ar

Resumen. El trabajo presenta experiencias en el modelado de datos sobre la dinámica vehicular en zonas urbanas. A partir de la investigación sobre software que permita realizar modelos de tránsito, surge luego la propuesta de crear nuevas estructuras de datos, modelos para simulación y un software que permita adecuarse a las necesidades particulares de investigación. Se pretende a través del presente estudio proponer metodologías para la captura y procesamiento de datos, no solo proveyendo formato sobre los datos obtenidos del modelado y posteriores simulaciones, sino que el aporte se realice a la luz de una experiencia de la computación aplicada al análisis y diseño de productos y datos útiles para la ingeniería de tránsito. La aplicación de sistemas GPS proporciona la automatización requerida para la obtención de datos masiva. Complementa el trabajo el modelado de datos para un adecuado almacenamiento.

Palabras Clave: Flujo Vehicular, Modelos de Tránsito, Congestión Vehicular.

1 Introducción

El proyecto en el cual se encuadra el trabajo tiene como objeto describir las características de la dinámica del flujo vehicular en ambientes urbanos, para la elaboración de una metodología estándar que automatice la captura, procesamiento y almacenamiento de datos, identificando las variables que determinan situaciones particulares, como por ejemplo la congestión vehicular [1]. Se busca reemplazar metodologías clásicas de aforo vehicular, donde generalmente se utilizan hojas de datos para asentar mediciones sobre velocidad, densidad, etc. La utilización de técnicas alternativas, como por ejemplo la del auto flotante [2], representa un esfuerzo para explicar inestabilidades del tránsito en función de la respuesta del conductor. Al utilizar un vehículo testigo circulando por las áreas de estudio, es posible obtener mediciones directamente desde el mismo. Tales mediciones

corresponden a la obtención de velocidades puntuales, que son utilizadas en etapas posteriores para establecer condiciones de congestión vehicular.
En la sección 2 se mencionan estudios preliminares que se encuadran en el tema de estudio. La sección 3 describe el diseño físico de la arquitectura computacional utilizada para el almacenamiento y cómputo de los datos recolectados: sus bases de datos, y una descripción de los algoritmos utilizados. En la sección 4 se menciona el modelo de datos propuesto, donde las características estructurales de las tablas de datos resultan de suma importancia para mantener un rápido acceso a los mismos. La sección 5 hace referencia a uno de los productos considerados por el grupo de trabajo: algoritmos para establecer rutas más cortas. Finalmente se presentan las conclusiones sobre el análisis, diseño, desarrollo y uso del sistema de información planteado.

2 Estado del Arte

2.1 Productos para la Asistencia al Modelado de Tránsito

Los estudios sobre tránsito y transporte [3][4], requieren del modelado y simulación de escenarios de circulación; lo cual requiere una correcta determinación de aquellos indicadores que permitan describir en forma clara y precisa situaciones y proyecciones de la dinámica vehicular. Se apunta a ofrecer soluciones alternativas sobre el procesamiento de datos provenientes de dispositivos GPS, atendiendo a la estructura de datos y algoritmos considerados para el planteo de escenarios particulares de flujo vehicular, determinación de indicadores de congestión, etc. Se indagó sobre el mercado de software destinado al modelado vehicular [5], encontrando que existen variadas metodologías para muestreo vehicular (por ejemplo, el método del auto flotante entre otros), y las correspondientes herramientas para simulación de escenarios de tránsito. Se han analizado distintas perspectivas que presentan estos productos para el análisis a través de la generación de micro y macro modelos. Dentro de las herramientas que se encuentran disponibles actualmente en el mercado, se destaca el software AIMSUN [6], el cual ha sido tomado como modelo debido al grado de generalidad que presenta. Las capacidades provistas por el software se centran principalmente en la realización de micro y macro modelos de una red de tránsito, como así también la simulación de maniobras reales. No obstante ello, el software no ha permitido avanzar sobre puntos de estudio u observación claves dentro de situaciones planteadas, por ejemplo: determinación de la congestión vehicular a través de datos históricos. Este estudio resulta sumamente conveniente para establecer políticas de circulación o diseño de obras civiles.

2.2 Estudios de Campo

En primer lugar se han utilizado hojas de datos para la recolección de datos sobre velocidades y densidad vehicular, los cuales permiten en una primera instancia analizar las posibilidades de automatización de las muestras, el análisis de

dispositivos de captura, etc. La experiencia resulta de utilidad para observar tanto las principales variables de medición (por ejemplo: velocidad media por tramo), como así también situaciones de circulación irregulares o anómalas. A partir de los estudios mencionados, se decide la implementación de herramientas propias para el almacenamiento, procesamiento y generación de información sobre los diferentes escenarios modelados.

3 Arquitectura Computacional de la Solución Propuesta

3.1 El dispositivo de Captura de Datos

Para la obtención de trazas se utiliza un dispositivo *GPS GARMIN Etrex Vista HCx* [7], el cual resulta adecuado para obtener datos que se ajusten tanto a las variables de medición requeridas para el posterior modelado, como así también al tiempo mínimo requerido entre registros (por ejemplo: cada un segundo). Por otro lado se analizó la aplicación de *SmartPhones* para la adquisición masiva de datos, ya que por su popularidad, es posible distribuir la toma de datos a lo largo de mayor cantidad de vehículos de prueba. Se considera por otro lado que estos dispositivos son extremadamente dependientes del software que poseen, no encontrando estándares que permitan una obtención de datos homogéneamente confiables en el corto plazo.

3.2 Diseño de la Aplicación

Si bien el sistema de información planteado no cuenta con las características de un sistema de procesamiento en tiempo de real, se pretende automatizar la captura del tren de datos provenientes del dispositivo GPS, de tal forma que resulte conveniente para la generación de información estadística. La figura 1 muestra en forma esquemática los diferentes subsistemas considerados para la obtención de información estadística.

Fig. 1. El sistema de información propuesto tiene por objeto el almacenamiento, procesamiento de datos y la posterior generación estadística sobre los escenarios modelados.

En primer lugar se cuenta con un conjunto de funciones que permiten el reconocimiento de diferentes tipos de trazas (*Parser*). Luego, a partir de la depuración de las trazas capturadas por los dispositivos, los datos son filtrados, eliminando aquellas muestras que no son útiles para el posterior procesamiento

(*Filtros* y *Conversion*). A continuación se trabaja sobre algoritmos que procesan los datos filtrados para ser almacenados en base de datos. Finalmente, se requiere del procesamiento de la información para la generación de escenarios posibles de circulación. Esto plantea el problema de tener que procesar grandes volúmenes de datos. El producto obtenido consiste en un conjunto de algoritmos, bases de datos normalizada y metodologías de procesamiento, que sirve como insumo para ser utilizado en futuros proyectos de modelado numérico de situaciones reales de tránsito. En resumen: el aporte esperado es la confección de una metodología de procesamiento masivo de datos de la dinámica vehicular, lo cual permita obtener información útil para crear nuevos modelos urbanos. Un avance del trabajo realizado se menciona en la sección 4.

3.3 Producción de la Información

La capa de *Producción* (figura 1) corresponde a un conjunto de algoritmos que establecen *a priori* información útil para la circulación de un vehículo sobre el área de estudio. Los algoritmos se orientan a la sumarización, ordenamiento, agrupamiento, etc, de conjuntos de datos provenientes de las trazas, y de acuerdo a metodologías estándares de muestreo establecidas por la Ingeniería de Tránsito. Las muestras correspondientes a los conjuntos de datos en bruto, provenientes de numerosas muestras obtenidas a través de los dispositivos de georeferenciación, son sistematizadas a fin de generar información para elaborar escenarios de dinámica vehicular. El procesamiento principal corresponde a la obtención de caminos más cortos entre dos puntos específicos del área de estudio, contando con información estadística sobre la circulación del vehículo de prueba sobre dicha área.

3.4 Almacenamiento y Clasificación

El almacenamiento principal se encuentra en una primera instancia sobre un conjunto de archivos de texto obtenidos del dispositivo. Se ha propuesto una metodología estándar de clasificación de tales ficheros para mantener un orden de acuerdo al vehículo testigo que toma la muestra, y otras variables que permiten describir el entorno en el cual se tomo la muestra, por ej: Característica del Vehículo Testigo, Día de la Semana / Fecha, Hora, Datos climáticos, etc.

Cada fichero representa una circulación particular del vehículo de prueba; luego, el conjunto de ficheros es introducido en la base de datos con el objeto de generar un único dominio de datos del cual poder generar información estadística. Dado la necesidad de almacenamiento masiva, y con el objeto de mantener almacenadas la totalidad de las muestras obtenidas del conteo vehicular, se requiere el desarrollo de un modelo de datos eficiente, para ser implementado con un motor de base de datos que resulte adecuado para la gestión de los datos. La confección de la base de datos consta de dos etapas:

1. **Información Estática**. Comprende la georeferenciación de puntos de interés a lo largo de las zonas de estudio. Estos puntos luego conforman una malla

interconectada. La información estática está conformada por tablas que contienen básicamente la información sobre esquinas, tramos de calles (por ejemplo, en la figura 2 se muestra un tramo en azul, cuyos datos representan las coordenada entre las cuales está comprendido, su nombre, etc.) y zonas de estudio; sus nombres y datos principales para identificarlas.

Fig. 2. La base de datos estática contiene información sobre coordenadas de puntos de interés e información sobre tramos.

2. **Información Dinámica**. Corresponde a la información obtenida a través de la captura permanente de datos. Está compuesta por velocidades promedio, máximas, mínimas por tramo, conjunto de tramos etc. La contrastación de esta información con la malla obtenida en la etapa 1 permite generar información estadística sobre la dinámica vehicular, observar zonas de congestión, etc.

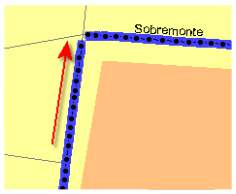


Fig. 3. La base de datos dinámica contiene información sobre recorridos de los vehículos testigos (velocidades y características de circulación).

Como se especifica en la sección 4, los datos almacenados en la base de datos son útiles para generar en tiempo de ejecución información útil para representar diferentes escenarios de circulación, por ejemplo a través de la constitución de matrices de adyacencia que expongan el paso entre diferentes puntos de la malla. El recorrido continuo del vehículo por la zona de prueba (ver figura 3) permite establecer el sentido de circulación entre diferentes tramos.

(a)

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b)

Fig. 4. El sentido de circulación de cada tramo permite establecer un grafo dirigido (a), y a continuación una matriz de adyacencia que lo contenga (b).

Para el ejemplo de la figura 4 se han tomado solamente 6 tramos de estudio (fig. 4a), lo que deriva en una matriz de adyacencia (fig. 4b) de 36 coeficientes con solamente 5 elementos no nulos. Esta información es generada por el vehículo testigo en forma dinámica a medida que circula a través de los puntos (esquinas) almacenados previamente en la base de datos estática.

3.5 Algoritmos de Filtrado de Datos

Se desarrollan algoritmos para ajustar los datos de acuerdo a las necesidades de información, como por ejemplo la existencia de situaciones de circulación vehicular anómalas, fallas del dispositivo de captura de datos, etc. Por ejemplo, en una captura de datos es común encontrar numerosas muestras con *vel=0* (velocidad igual a cero, o auto detenido), o la captura de datos donde el vehículo testigo realiza diferentes maniobras. Estos eventos resultan inútiles para establecer información de interés, como por ejemplo la velocidad de marcha.

3.6 Reconocimiento de trazas heterogéneas

Estos algoritmos están destinados a la conversión de unidades (por ej. coordenadas con diferente formato) y eliminación de datos inútiles. A partir de los diferentes formatos obtenidos, resulta útil establecer una metodología de reconocimiento. Para ello se genera un compilador capaz de reconocer diferentes formatos de traza, proveniente de diferentes dispositivos. La figura 5 ejemplifica un formato básico de traza donde se observan entre otros datos, la coordenada y la velocidad puntual.

```
...
75 05/07/2013 18:30:49  0 m 0:00:01        0 km/h     S32.89669 W68.85365
76 05/07/2013 18:30:50  0 m 0:00:01        0 km/h     S32.89669 W68.85365
77 05/07/2013 18:30:51  8 m 0:00:01       29 km/h     S32.89669 W68.85365
78 05/07/2013 18:30:52 1 m 0:00:01         5 km/h     S32.89662 W68.85364
79 05/07/2013 18:30:53 2 m 0:00:01         6 km/h     S32.89661 W68.85364
80 05/07/2013 18:30:54 2 m 0:00:01         6 km/h     S32.89659 W68.85364
81 05/07/2013 18:30:55 2 m 0:00:01         8 km/h     S32.89658 W68.85363
82 05/07/2013 18:30:56 3 m 0:00:01        10 km/h     S32.89656 W68.85363
83 05/07/2013 18:30:57 3 m 0:00:01        12 km/h     S32.89653 W68.85363
84 05/07/2013 18:30:58 3 m 0:00:01        12 km/h     S32.89650 W68.85362
85 05/07/2013 18:30:59 3 m 0:00:01        11 km/h     S32.89647 W68.85362
86 05/07/2013 18:31:00 3 m 0:00:01        10 km/h     S32.89645 W68.85361
87 05/07/2013 18:31:01 3 m 0:00:01        10 km/h     S32.89642 W68.85361
...
```

Fig. 5. Ejemplo de una traza de datos. Cada línea representa un registro obtenido por el dispositivo de captura de datos.

El compilador de trazas [8], es desarrollado a través de las herramientas Flex y Bison [9] [10], y comprende las siguientes tareas: 1) Incorporación y lectura del fichero de trazas, 2) Determinación de errores en las mediciones y 3) Compilación de la traza hacia un formato estandarizado.

La consolidación de datos a partir de trazas heterogéneas permite luego obtener una mayor cantidad de muestras, contando con una base de datos con trazas provenientes de diferentes dispositivos o diferentes modalidades de captura.

4 Modelo de Datos

4.1 Adquisición de Datos para la base de Datos Estática

Se han realizado algoritmos que permiten a través de los datos obtenidos en crudo, establecer la existencia de esquinas (o puntos de interés), con lo cual es posible alimentar la base de datos en forma continua, y tras la circulación permanente del vehículo de prueba por zonas de estudio.

4.2 Adquisición de Datos para la base de Datos Dinámica

La base de datos con información dinámica contiene información recolectada a través de la técnica del auto flotante. La información obtenida se puede caracterizar en base a: 1) Información sobre el paso de un tramo a otro, 2) Información sobre la velocidad puntual en coordenadas específicas.

Se obtienen permanentemente datos a través del dispositivo GPS, los cuales alimentan la base de datos de *esquinas* y *tramos*. En la figura 6 se observa que a medida que se obtienen más trazas a través de la circulación libre del vehículo a través de la ciudad, se registran más datos *paso(tramo1,tramo2)*. A simple vista se observa mayor densidad en la matriz de adyacencia.

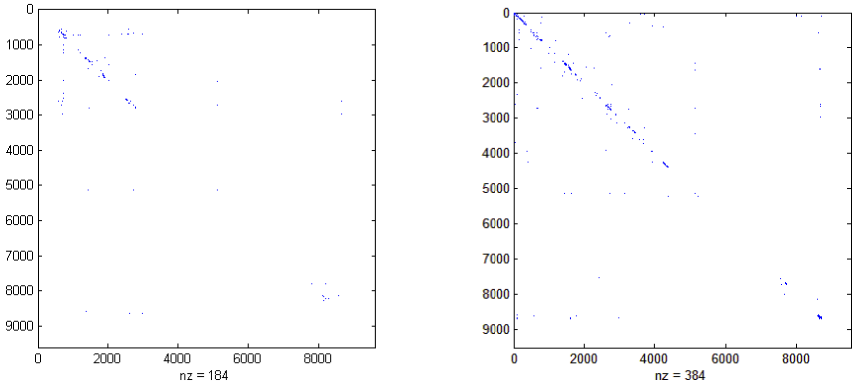


Fig. 6. La Matriz de Adyacencia presenta coeficientes que determinan el paso de un tramo *paso(fila,columna)*. (a) y (b) son dos momentos diferentes del estado de la base de datos, donde la segunda matriz presenta mayor cantidad de entradas.

Se puede visualizar en la figura que la densidad de la matriz se centra a lo largo de la diagonal principal, situación ventajosa al trabajar la matriz numéricamente.

4.3 Formato de Representación de la matriz de adyacencia

A partir del modelado de datos se cuenta con una matriz binaria de 81.000.000 coeficientes, donde una entrada diferente de 0 en (*fila,columna*) determina el paso entre (*tramox,tramoy)* donde *tramox=fila* y *tramoy=columna* del coeficiente. Esta metodología permite describir un grafo dirigido, donde los nodos representan a cada tramo de estudio, y las aristas el sentido de circulación vehicular entre dichos tramos.

Al almacenar la matriz completa en memoria RAM se requieren *9.000*9.000*2 =154,4952Mb* (para un coeficiente que ocupa 2 bytes en memoria), siendo el porcentaje ocupación del $4{,}74e^{-4}$ %.

Al operar con la matriz de adyacencia (como se especifica en la siguiente sección) se torna necesario disponer de arreglos auxiliares, con lo cual, las magnitudes expresadas crecen considerablemente.

Se debe diferenciar el almacenamiento temporal de la matriz en memoria RAM, con la representación de la misma en disco. En este último se almacena la matriz completa con el objeto de mantener una mejor reutilización de datos entre los programas que utilizan tales ficheros. Por otro lado, se busca una representación adecuada de la matriz en memoria RAM, teniendo en cuenta las características de dispersión de la misma. La representación de tipo *coordenada* , como un tipo básico para la representación de una matriz rala [11] lleva a almacenar únicamente los elementos no nulos junto con datos que determinan su ubicación *(fila,columna)*. Este almacenamiento permite reducir la cantidad de espacio en memoria principal a menos de 3Kb para la matriz del ejemplo. Por otro lado, la complejidad algorítmica de los métodos numéricos que operan sobre ella aumenta al requerir el mapeo de un coeficiente sobre una nueva estructura de datos diferente a un arreglo bidimensional.

5 Algoritmos para Encontrar Trayectorias.

5.1 Algoritmos para encontrar rutas

Muchos estudios dentro de las ciencias de la computación hacen referencia al esquema de computación planteado, conociéndose en la literatura como el problema APSP (*All Pairs Shortest Path*) aquel que tiende a encontrar caminos más cortos (o favorables) entre dos puntos en un grafo. A partir de la matriz de adyacencia es factible plantear la solución a problemas de este tipo. En este trabajo se destaca una situación especial al presentarse una matriz de adyacencia de grandes dimensiones. Luego, dicho problema, que en la comunidad de ciencias de la computación presenta un constante desafío de resolución al intentar minimizar tiempos de resolución, se incrementa al contar con un grafo dirigido con una cantidad de componentes considerable.

Se propone en primer lugar una solución que lista y aisla desde la base de datos aquellas relaciones que permiten establecer el camino entre dos puntos particulares en la zona de estudio, podando el resto del grafo dirigido.

La potenciación de la matriz de adyacencia representa algunos de los cálculos más complejos desde el punto de vista computacional al consumir considerables recursos y tiempo de ejecución. Este cómputo es realizado una considerable cantidad de veces al describir el camino más ventajoso entre dos puntos. Se realiza la potenciación de la matriz como una sucesión de productos, donde

$$P=A^2=A*A. \quad (\mathbf{1})$$

representa la matriz de adyacencia para pasos de longitud 2 entre dos puntos de interés. Se recuerda que la complejidad algorítmica de dicho producto es $O(n^3)$ para n cantidad de coeficientes. De acuerdo a la cantidad de coeficientes, para un almacenamiento convencional como el descrito en la sección 4.3 se prevé que la carga computacional es considerablemente alta para el cálculo de (1).

5.2 Procesamiento Paralelo de la Información

Con el objeto de paliar los altos tiempos de cómputo mencionados en 5.1, se propone la paralelización del proceso principal. Se procede a la paralelización del algoritmo de multiplicación sucesiva de matrices y se proponen dos algoritmos de multiplicación. En primer lugar, un algoritmo basado en memoria distribuida y paso de mensajes entre nodos para ser ejecutado en un *cluster* de computadoras. Esta solución se materializa a través de la utilización de la librería MPI [12].

Luego se propone un algoritmo con memoria compartida para ser ejecutado en una arquitectura superescalar multinúcleo, a través de la paralelización de hilos *Posix* [13].

En ambos casos se consigue obtener un *speedup* considerable frente a la solución secuencial.

Conclusiones y Trabajo Futuro

El trabajo permitió establecer un sistema de información que abarca desde la captura de datos, hasta la generación de estadísticas en torno a información real de circulación de vehículos testigo. Dicha información revela situaciones particulares, que son difíciles de obtener a través del aforo vehicular clásico.
Se facilita a investigadores e ingenieros relacionados con el tránsito y transporte el modelado de escenarios de flujo vehicular, proveyendo una estructura y conjunto de datos de fácil almacenamiento, acceso y manipulación; posibilitando a los mismos acceder a datos procesados o en crudo: condiciones no prevista en la mayoría de los software de modelado de tráfico vehicular de mayor incidencia en el mercado.
Para aquellos procesos que demandan mayor cantidad de tiempo de cálculo se proveyó de algoritmos paralelos, con los cuales se consiguieren considerables mejoras en los tiempos de ejecución.
Como trabajo futuro, el grupo analizará tecnologías para el procesamiento en tiempo real del tren de datos como resulta el GPRS. También se prevé como trabajo futuro la cuantificación y documentación de la ganancia en tiempo de ejecución experimentada con las nuevas arquitecturas de procesamiento paralela.

Referencias

1. Thompson I., Bull, A.: La Congestión del Tránsito Urbano: Causas y Consecuencias Económicas y Sociales. CEPAL (Naciones Unidas, División de Recursos Naturales e Infraestructura, Unidad de Transporte). Serie Recursos Naturales e Infraestructura (25). Chile, (2001). ISBN: 92-1-321865-6
2. Hall R.W.: Handbook of Transportation Science. 2º Edition. Edited by Randolph W. Hall, University of Southern California. Kluwer's Internation Series. USA, (2003). ISBN: 1-4020-7246-5.
3. Cal, R., Reyes, M.: Ingeniería de Tránsito. Fundamentos y Aplicaciones 7ª Edición. Alfaomega. (1995). ISBN: 970-15-0109-8.
4. Schöbel, A.: Optimization in Public Transportation: Stop Location, Delay Management and Tariff Design in a Public Transportation Network. Georg-August University (Göttingen-Alemania). Springer. USA, (2006). ISBN: 978-0-387-32896-6.
5. Software para el Modelado de Tránsito. Julio Monetti. Disponible en http://grid2.frm.utn.edu.ar/proyectotransito.
6. TSS. AIMSUN. Disponible en http://www.mctsoft.com/html/fr_mct.htm.
7. GARMIN. https://buy.garmin.com/en-GB/GB/prod8703.html
8. Compilador de Ficheros de Trazas. J. Monetti. http://grid2.frm.utn.edu.ar/proyecto.
9. Aaby, A.: Compiler Construction using Flex and Bison.. Walla Walla College. (2005). Disponible en aabyan@wwc.edu.
10. Levine, J.: flex & bison 1º Edición. O'Reilly Media. USA, (2009). ISBN: 978-0596155971
11. Tewarson R.: Sparse Matrices.. Department of Applied Mathematics and Statistics. Academic Press Inc. USA, (1973). ISBN: 9780126856507.
12. Gropp, W., Lusk E. Thakur, R.: Using MPI-2. Advanced Features of the Message-Passing Interface. The MIT Press. Cambridge – Massachusetts. London, England. (1999). IBN: 0-262-057133-1.
13. Butenhof, D.: Programming with Posix Threads. Addison Wesley Longman Inc. USA, (1997). ISBN: 0-201-63392-2.

Planeamiento Estratégico para Compartir Información en la Administración Pública[1]

Ignacio Marcovecchio[1], Elsa Estevez[2], Pablo Fillottrani[1,3],

[1] Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur,
Av. Alem 1253 – Bahía Blanca, Argentina
[2] United Nations University – IIST, Center for Electronic Governance,
P.O. Box 3058, Macao SAR, China
[3] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina

ignaciomarcovecchio@gmail.com, elsa@iist.unu.edu, prf@cs.uns.edu.ar

Abstract. Poder compartir información es fundamental para el funcionamiento efectivo y eficiente de la administración pública, y en especial, para iniciativas de gobierno electrónico. Tal es su importancia que el nivel de madurez más alto en gobierno electrónico se alcanza cuando las agencias de gobierno son capaces de compartir información. Sin embargo, la implementación de iniciativas basadas en compartir información (CI) resulta difícil como consecuencia de las barreras que suelen existir. Para poder atacar una problemática tan amplia y compleja, todas las iniciativas deben estar coordinadas y deben contribuir a alcanzar los mismos objetivos. Por este motivo, es necesario que el gobierno cuente con una estrategia para encarar organizadamente las acciones relacionadas con CI. En este trabajo se presenta un proceso que tiene por objetivo asistir a los responsables de gobierno en la definición de planes estratégicos para compartir información en el sector público. El proceso consta de ocho pasos que guían el desarrollo de la planificación estratégica que apuntan a facilitar la planificación y lograr buenos resultados al CI.

Keywords: Compartir Información; Planeamiento Estratégico; Chief Information Officer (CIO) de Gobierno (GCIO)

1 Introducción

Compartir información en el ámbito de gobierno se define como el conjunto de todas las actividades que tienen por objetivo obtener, mantener, usar y proteger información en el sector público [1]. Poder compartir información – tanto dentro de una agencia de gobierno, como entre agencias o entre gobiernos – es fundamental para el funcionamiento de la administración pública, y en especial, para iniciativas de gobierno electrónico. Tal es su importancia que el nivel de madurez más alto en gobierno electrónico se alcanza cuando las agencias de gobierno son capaces de

[1] Este trabajo de investigación fue financiado parcialmente por la Fundación Macao, Macao SAR, mediante la ejecución de uno de los proyecto del Programa e-Macao.

compartir información. Sin embargo, y pese a su importancia, la implementación de iniciativas de CI suele resultar difícil de conseguir como consecuencia de las barreras que suelen existir: barreras tecnológicas, culturales, organizacionales, políticas, etc.

Debido a su importancia y a su naturaleza transversal – ya que incluye a distintas entidades del gobierno – las iniciativas de CI requieren un alto grado de coordinación, y son usualmente asignadas como responsabilidad de la oficina del Chief Information Officer de Gobierno (GCIO). Como responsable del liderazgo y la coordinación de todas las iniciativas del gobierno en materia de tecnología de la información y las comunicaciones (TICs), el GCIO juega un rol central y tiene un especial interés en maximizar el compartimiento de la información.

Para poder atacar una problemática tan amplia y compleja como la de CI, las iniciativas deben estar coordinadas y deben contribuir a alcanzar los mismos objetivos. Por este motivo, es necesario que el gobierno cuente con una estrategia para encarar organizadamente las acciones relacionadas con CI. Los GCIOs disponen de herramientas para la definición de estrategias; una de estas herramientas es el planeamiento estratégico [13]. Como la elaboración de planes estratégicos es una tarea compleja y que involucra a muchos actores, existen procesos que guían el desarrollo de los mismos. Estos procesos, sin embargo, no son específicos para cada problemática, sino que se adaptan a cualquier tipo de organización y se utilizan para planificar estrategias de distintos dominios. Para asistir a los GCIOs en la definición de planes estratégicos para CI, en este trabajo se seleccionó un proceso para el planeamiento estratégico de cuestiones relacionadas con TICs y se lo ajustó y personalizó para el dominio específico de CI. Con esta herramienta, los gobiernos no tendrán que repetir estas acciones cuando decidan planificar sus estrategias de CI, sino que podrán utilizar y perfeccionar el proceso existente.

El presente trabajo tiene por objetivo asistir a los líderes de tecnología del sector público en la planificación estratégica para la implementación y el fortalecimiento del uso compartido de la información en la administración pública. Para lograr este objetivo se propone esencialmente una herramienta – un proceso para la definición de los planes estratégicos para CI; que sirve para guiar a las partes involucradas durante el proceso de planificación de iniciativas dedicadas a CI.

El resto del trabajo está organizado de la siguiente manera: la Sección 2 presenta y describe el proceso de planeamiento estratégico, ajustado para el dominio de CI; la Sección 3 presenta trabajos relacionados. Finalmente, la Sección 4 presenta las conclusiones obtenidas y las líneas de investigación planeadas a futuro.

2 Proceso para Realizar un Plan Estratégico para CI

El proceso que se define a continuación para la definición de un plan estratégico de CI se basa en el proceso de planeamiento estratégico para gobernanza electrónica descripto en [2]. Este proceso se compone de ocho pasos (ver Fig. 1), los cuales se ajustaron con las modificaciones necesarias para su utilización en el contexto de CI para gobierno electrónico. A continuación se describen en detalle cada uno de los pasos del proceso resultante.

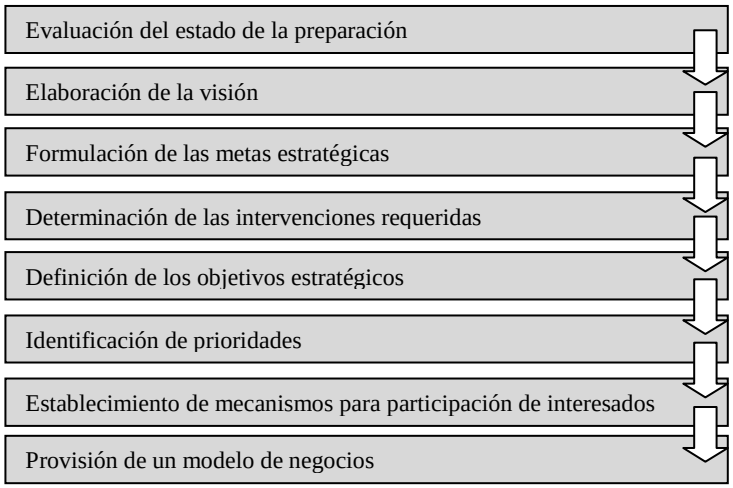


Fig. 1. Pasos del proceso para planeamiento estratégico para CI.

2.1 Evaluación del Estado de Preparación

El objetivo de evaluar el estado de preparación de la administración pública para compartir información es obtener la información necesaria para la definición del plan estratégico. Una acertada evaluación del estado de preparación es fundamental para la correcta definición de los elementos del plan. Además, la información obtenida en este paso resulta de gran utilidad para medir los resultados del plan.

El estado de preparación del gobierno con respecto a la capacidad para CI se puede determinar siguiendo metodologías como las presentadas en [3] y [4]. Esta metodología se basa en obtener el conocimiento sobre el estado general del arte de la administración pública, la situación actual con respecto al uso de la información y las percepciones de los principales interesados con respecto a las prácticas para compartir información. La metodología está conformada por cuatro componentes: i) un proceso para guiar el ejercicio de evaluación, ii) un modelo conceptual que ayuda en la definición de las áreas a tener en cuenta en la evaluación, iii) los instrumentos a utilizar para la evaluación y, iv) guías prácticas para la ejecución de cada una de las actividades.

Algunas de las áreas a considerar en una evaluación que ofrezca una visión completa del estado de preparación son:

- *Preparación Política* – mejorar el uso de información requiere fuerte compromiso por parte de los líderes políticos. Para esto debe existir compromiso político para mejorar el uso de la información, conocimiento de la importancia y los beneficios de mejorar el uso y la forma en la que se comparte la información, y liderazgo capaz de administrar los cambios necesarios. Todos estos aspectos deben ser relevados, junto con las percepciones de los referentes del nivel político.
- *Preparación Regulatoria* – un marco regulatorio bien definido es necesario para asegurar que se comparta información tanto dentro del gobierno, como así también entre el gobierno, los ciudadanos y las organizaciones. Este marco es además

necesario para garantizar las condiciones económicas para el acceso a la infraestructura tecnológica, a los servicios y al equipamiento necesario. Por lo tanto, se debe estudiar la legislación existente con respecto a privacidad, utilización de estándares, el grado de liberalización de la industria de las telecomunicaciones y el ambiente fiscal para la adquisición de equipamiento de TICs.

- *Preparación Organizacional* – la implementación o la mejora de la forma en la que se comparte información suele requerir e implicar cambios organizacionales en la estructura del gobierno. Para justificar estas transformaciones es necesario disponer de información sobre los flujos de información existentes, los procesos que se llevan a cabo para ofrecer servicios, los datos solicitados por cada agencia del gobierno, las relaciones entre los distintos organismos del gobierno, los niveles de autoridad y control entre estos organismos, las experiencias de colaboración en el pasado, la existencia de unidades centrales de coordinación, etc.
- *Preparación Cultural* – los aspectos culturales suelen generar resistencia a los cambios. Cuanto más se sepa sobre estos aspectos culturales, mejor se pueden planificar las estrategias para afrontar aquellos que puedan resultar negativos. Entre los aspectos a considerar se encuentran los niveles de educación, la actitud hacia los cambios, la cultura para compartir información y conocimiento, el nivel de educación en tecnología de la información y las comunicaciones, la cultura organizacional en la administración pública, la orientación de servicio hacia los clientes por parte de los servidores públicos, etc.
- *Preparación Económica* – los costos para poner en marcha un sistema de intercambio de información suelen ser altos. Por este motivo, se debe tener una noción precisa de los recursos disponibles en la administración pública para poder realizar una planificación adecuada de su utilización. Además, se deben estudiar los posibles mecanismos de financiamiento para hacer frente no sólo a la puesta en marcha sino también al sostenimiento en el tiempo del sistema. Algunos de los aspectos que se deben estudiar en esta área son los recursos financieros disponibles, la estructura de ingresos del gobierno y el acceso a mecanismos de financiamiento alternativos, la posibilidad de cooperación con el sector privado, el acceso a los mercados de capitales, etc.
- *Preparación en Infraestructura* – una infraestructura tecnológica deficiente representa un obstáculo para implementar y sostener un sistema para el intercambio de información. Por este motivo se deben relevar los recursos tecnológicos existentes y se deben evaluar los factores que puedan resultar problemáticos, para poder planificar la forma de superarlos. Fundamentalmente se debe conocer la disponibilidad de recursos de hardware, software y sistemas de telecomunicaciones, así como también la existencia de sistemas legados y las condiciones demográficas y geográficas que pueden afectar la distribución económica, y por lo tanto, influir en la disponibilidad de infraestructura (tanto pública como privada).
- *Preparación de Datos e Información* – para lograr que la información pueda ser compartida electrónicamente deben existir sistemas de información, bases de datos y procesos de trabajo que soporten el funcionamiento del gobierno. Los datos a considerar cuando se evalúa esta área incluyen la disponibilidad y accesibilidad a los datos, los procedimientos para la captura y la estandarización de datos e información, la calidad de los datos y los mecanismos de seguridad para su acceso, la capacidad para analizar los datos y utilizar la información, entre otros aspectos.

2.2 Elaboración de la Visión

En este paso se busca definir un futuro estado que el gobierno desea alcanzar a largo plazo en materia de CI. El instrumento que se utiliza para plantear estas ideas se llama visión y su declaración sirve de guía general para la implementación del proceso. Debido a su importancia, existen procesos para la construcción de la visión. Un ejemplo es el proceso presentado en [2], que involucra cinco pasos. A continuación se presentan estos pasos y se los explica en el contexto de IS.

1) *Identificar y consultar a los interesados* – lograr consenso y apoyo en las iniciativas requiere involucrar y contar con el apoyo de los sectores involucrados.
2) *Permitir a los interesados presentar y explicar sus propias visiones* – para fomentar la participación y el sentido de pertenencia, además de contemplar intereses y preocupaciones de las partes involucradas, se deben contemplar mecanismos para que los interesados puedan exponer sus propias visiones.
3) *Desarrollar una versión preliminar de la visión* – esta propuesta se debe desarrollar en base a las visiones de los interesados.
4) *Alinear la visión con las necesidades y oportunidades generales* – la visión de CI en el gobierno tiene que estar alineada con la visión general del gobierno con respecto a las TICs. Debido a la naturaleza transversal de CI, es conveniente que la visión sea definida por el más alto nivel de administración pública (i.e. nacional).
5) *Consolidar y consensuar una versión final de la visión* – para ser efectiva, la versión final debe respetar las siguientes características: i) debe ser clara, intuitiva y simple; ii) debe establecer claramente el alcance (establecer qué será hecho y qué no será hecho); iii) debe considerar necesidades y oportunidades; iv) debe contar con el consenso de los interesados; y v) debe estar alineada con la estrategia nacional de TICs.

2.3 Formulación de las metas estratégicas

Las metas estratégicas son declaraciones que establecen la dirección de CI y que se basan en lo postulado por la visión. Su objetivo es definir los logros que se pretenden alcanzar a largo plazo. Las metas cumplen tres funciones principales: 1) establecen el estado futuro deseado que la organización quiere alcanzar en materia de CI, por lo que constituyen principios generales que deben ser seguidos por los miembros de la organización; 2) proporcionan una lógica o razón fundamental para la existencia de las iniciativas de CI; 3) proporcionan un conjunto de estándares con los que se puede contrastar el rendimiento de las acciones tomadas.

Algunos temas que las metas consideran incluyen: 1) efectos sociales y económicos provocados por el desarrollo de CI; 2) ofrecimiento de servicios públicos de calidad; 3) impacto a partir de la posibilidad de realizar mejores controles, las mejoras en la eficiencia y la efectividad, y la reducción de costos; 4) impacto en la gobernabilidad favorecido por las mejoras en los procesos de toma de decisiones.

Una vez definidas tanto la visión como las metas estratégicas es recomendable identificar cuáles son los mayores desafíos y barreras para concretarlas. Tener estos

aspectos presentes durante la implementación del proceso contribuye a prevenir y a reaccionar de mejor manera ante los obstáculos que se puedan presentar.

2.4 Determinación de las intervenciones requeridas

Este paso consiste en identificar los aspectos necesarios para la creación de un ambiente que favorezca el desarrollo de iniciativas de CI. La información de entrada al proceso incluye los resultados del estudio del estado de preparación (sección 2.1), la visión estratégica (sección 2.2) y las metas estratégicas (sección 2.3). Las intervenciones requeridas pueden considerar, entre otras, las siguientes dimensiones:

- *Intervenciones legales* – nuevas leyes y normas pueden ser necesarias para la adopción de CI. Algunos cuestiones legales a considerar incluyen: la integración y el intercambio de datos entre agencias públicas; el uso de la información pública por terceras partes, preservando la privacidad y la seguridad; el intercambio y las transacciones digitales entre agencias de gobierno, ciudadanos y empresas; el reconocimiento de los intercambios digitales de información y las transacciones digitales; etc. Ejemplos de medidas regulatorias tomadas por gobiernos líderes en gobierno electrónico son la Ley de Protección de Datos del Reino Unido [5] que "protege la privacidad personal y permite el libre flujo de datos personales por armonización", la ley de Libertad de Información Electrónica de los Estados Unidos [6] que, entre otras cosas, "instruye a todas las agencias federales a usar tecnologías de información electrónica para fomentar la disponibilidad pública de documentos electrónicos" y además "le otorga a los individuos el derecho de acceder a los registros que se encuentran en posesión del gobierno federal", y la ley de Firma Electrónica de la Unión Europea [7] que "reconoce las firmas electrónicas dentro de la Unión Europea y que pueden ser usadas como evidencia en procedimientos legales".
- *Intervenciones organizacionales* – una buena práctica consiste en la definición de una entidad central de coordinación. La coordinación puede ser realizada por una agencia, un equipo dentro de un ministerio o un equipo creado específicamente para tal fin. La principal responsabilidad de esta entidad consiste en coordinar la implementación de la estrategia de CI. Adicionalmente, es responsable de realizar revisiones periódicas sobre el estado de preparación, coordinar campañas de concientización, asistir en las posibles relaciones con el sector privado, etc. Adicionalmente, una de las responsabilidades esenciales a este equipo es la de monitorear, evaluar y reportar el avance logrado en materia de CI.
- *Intervenciones en recursos humanos* – todos los interesados deben ser preparados con los conocimientos necesarios para desenvolverse en un contexto de CI. El tipo de capacitación y la forma de impartirla deben planificarse y ofrecerse de acuerdo a las necesidades y características particulares de cada sector.
- *Intervenciones en comunicación* – comunicar los objetivos y los logros obtenidos contribuye a crear conciencia y fomentar la participación de los interesados. Las estrategias de comunicación tienen que tener por objetivo crear interés y generar expectativas sobre los beneficios de CI. Los destinatarios principales de las

campañas de comunicación deben incluir a políticos, tomadores de decisiones, empleados de la administración pública, empresas privadas y ciudadanos.

- *Intervenciones en tecnología* – la infraestructura tecnológica es fundamental para el desarrollo de cualquier iniciativa de gobierno electrónico, en particular, para todas aquellas relacionadas con CI. Las intervenciones que debe realizar el gobierno en materia tecnológica deben apuntar a contar con una red de telecomunicaciones confiable y con costos accesibles, desarrollar políticas nacionales de TICs, asociarse con el sector privado para resolver cuestiones técnicas, y tener acceso a mejores prácticas internacionales para superar las limitaciones tecnológicas.

2.5 Definición de los objetivos estratégicos

Los objetivos estratégicos son declaraciones específicas y medibles sobre las metas estratégicas. Estos objetivos deben cubrir las acciones específicas a llevar a cabo y una definición de tiempos para su cumplimiento. Según [8], los objetivos se encuentran bien definidos si respetan las siguientes premisas: 1) son alcanzables, 2) son compresibles, 3) están ubicados en un horizonte temporal, 4) se derivan de las metas estratégicas de la organización, 5) se pueden transformar en tareas específicas, 6) posibilitan la concentración de recursos y esfuerzos, 7) no incluyen abstracciones, y 8) deben poder ser cuantificados o expresados en cifras. Los objetivos se materializan en la práctica a través de la implementación de programas y proyectos.

2.6 Identificación de prioridades

La cultura de compartir información no se logra a través de una única iniciativa, sino que se debe lograr a través de pasos concretos sobre los cuales se pueda construir credibilidad y sirvan para acercarse a las metas definidas. No existe una única regla que establezca cómo priorizar los objetivos, pero existen distintos criterios que pueden tenerse en cuenta al momento de establecer prioridades. Algunos de estos criterios a considerar incluyen: i) los recursos disponibles, ii) la sostenibilidad, y iii) la factibilidad de concretarlos – tanto desde el punto de vista tecnológico como institucional. La recomendación es no utilizar estos criterios de manera aislada, sino buscar formas de combinarlos, ponderando su importancia de acuerdo al contexto, a lo establecido por la visión y las metas estratégicas. La Fig. 2 propone una forma organizada para priorizar objetivos de acuerdo a criterios definidos.

	Criterios								Σ	Prioridad
Objetivos									74.8	5
									93.5	2
									87.6	3
									98.0	1
									66.5	6
									34.3	7
									86.8	4

Fig. 2. Matriz de prioridades.

El principio general para la priorización propone que se deben tener presentes los objetivos a largo plazo, pero que se deben concretar y consolidar los objetivos de corto plazo. Los criterios utilizados para la priorización deben combinarse con el impacto que producirán. El impacto se puede evaluar en distintas dimensiones, como por ejemplo: 1) *Impacto social* – nuevas oportunidades de empleo, mayor utilización de servicios, facilidad de operación, reducción de tiempos de resolución, mayor inclusión, etc.; 2) *Impacto económico* – reducción de costos, nuevas oportunidades de negocios, mejoras en los tiempos, etc.; y 3) *Impacto en la gobernabilidad* – mayor transparencia, combatir la corrupción, seguridad, privacidad, control, etc.

2.7 Establecimiento de mecanismos para participación de interesados

Existen diversos mecanismos para fomentar la participación de los interesados. Una forma de involucrarlos en el proceso de CI puede ser, por ejemplo, a través de un proceso de consulta pública. Un caso exitoso de inclusión de interesados es el resultado de la Política de Compromiso de los Interesados llevada adelante por el Departamento de Gobierno Local, Deportes y Recreación del gobierno de Queensland, Australia [9]. En ella se definieron seis principios para la participación de los interesados: inclusión, acercamiento, respeto mutuo, integridad, afirmar la diversidad y agregar valor.

Para poder determinar y asignar responsabilidades es necesario identificar a los interesados. Una posible clasificación de los roles de los interesados involucrados en las problemáticas de CI se presenta a continuación: 1) *Equipo* – personas van a trabajar directamente en los proyectos de CI; 2) *Proveedores* – proveedores de tecnologías, recursos y experiencias; 3) *Operadores* – empleados de las agencias que van a operar los sistemas de CI; 4) *Campeones* – entidades que conducen y buscan justificación para los proyectos; 5) *Patrocinadores* – entidades a cargo de los gastos de los proyectos; 6) *Propietario* – agencia que va poseer y usar los sistemas; y 7) *Otros* – recursos involucrados que tienen influencia significativa en el proyecto.

2.8 Provisión de un modelo de negocio

El modelo de negocio es un plan para asegurar la sostenibilidad de CI desde el punto de vista de su adopción y de los recursos necesarios. El modelo de negocio debe determinar cómo se van a desarrollar las soluciones de CI, las posibles opciones de financiamiento, los mecanismos para atraer la participación del sector privado, etc.

La disponibilidad de fondos determina el tipo de proyectos que se pueden llevar a cabo. Siendo CI una de las disciplinas principales para lograr el desarrollo de gobierno electrónico, un porcentaje de los fondos para CI puede provenir de los presupuestos definidos para tal propósito. Mientras que el financiamiento para los proyectos de CI puede provenir de los presupuestos de las agencias involucradas, una práctica que ha resultado exitosa para promover la colaboración entre distintos organismos de gobierno es la de disponer de presupuestos adicionales sólo para proyectos en los que participen colaborativamente más de un organismo. Otras estrategias de financiamiento incluyen variantes de las asociaciones publico-privadas.

3 Trabajos Relacionados

En la literatura existen varios trabajos relacionados con planeamiento estratégico de TI (PE) y con CI en gobierno. Con respecto a PE, [14] propone un plan para aplicar gobernanza de TI en el sector público basado en teorías de gerenciamiento participativo, planeamiento estratégico de TI, gobernanza de TI y administración pública. [15] presenta un modelo para evaluar el nivel de madurez de planeamiento estratégico de TI en organizaciones públicas de Brasil. En [16], se explica un enfoque, incluyendo un proceso para el planeamiento estratégico de TI para librerías e instituciones de educación superior. El proceso define las siguientes tareas: 1) definición del plan de proyecto con la alta gerencia, 2) realización de un análisis FODA (fortalezas, oportunidades, debilidades y amenazas), 3) evaluación del entorno y uso actual de TI en la librería, 4) evaluación de las TI disponibles en el mercado, 5) revisión del plan de acción con los interesados, 6) conducción de entrevistas con los principales interesados, 7) conducción de evaluaciones, 8) definición de un plan estratégico que incluye identificación, priorización y estimación de presupuesto para las iniciativas, 8) revisión del plan por parte del comité designado, 9) comunicación del plan y 10) revisiones periódicas del plan. Comparando las contribuciones de estos trabajos con los resultados presentados en este artículo, la más similar es [16]. La mayor diferencia es que ambos definen distintos niveles de granularidad para las tareas – el proceso propuesto para CI define tareas separadas para la definición de la visión, metas, intervenciones, objetivos y prioridades; mientras que el proceso para las librerías engloba todas estas tareas en una sola actividad (la número 8). Adicionalmente, el proceso aquí propuesto es específico para CI en el sector público.

Con respecto a trabajos relacionados con CI, existen varios que definen perspectivas de CI, los más relevantes incluyen [11] [12] [17]. Estos trabajos se utilizaron para definir las dimensiones de la evaluación de la preparación y los tipos de intervenciones requeridas.

4 Conclusiones

El nivel de madurez más alto en gobierno electrónico se alcanza cuando las agencias de gobierno son capaces de compartir información. Esto refleja la importancia que tiene y la complejidad que reviste lograr la administración eficiente de la información dentro de la administración pública.

Este trabajo propuso la utilización de herramientas del análisis estratégico como soporte para el análisis y la planificación de las acciones que deben llevar adelante los gobiernos para implementar y fomentar el uso compartido de la información en el ámbito de la administración pública. El proceso de planificación estratégica de CI apunta a servir de guía para que las decisiones y acciones que impactan en la información del sector público estén alineadas con los objetivos de utilizar y compartir la información de manera eficiente y eficaz, eliminado duplicidad y contribuyendo a lograr una mejor gobernabilidad.

La contribución de este trabajo es la definición de un proceso para la planificación estratégica de CI. El proceso cubre todo el ciclo de vida del ejercicio de planeamiento

estratégico y para cada una de las actividades se describen sus objetivos, la forma de llevarlos a cabo y los resultados esperados, acompañados de ejemplos reales.

Las líneas de trabajo futuro incluyen la validación del proceso propuesto mediante su instanciación en administraciones públicas de distintos niveles de gobierno (nacional, provincial y municipal), la definición de marcos de trabajo para la implementación de iniciativas de CI, y la construcción de herramientas automáticas que faciliten el proceso de planeamiento e implementación de tales iniciativas.

Bibliografía

1. Estevez, E., Fillottrani, P., Janowski, T., Ojo, A. *Government Information Sharing – A Framework for Policy Formulation*. In: Chen, Y.-C. and Chu, P.-Y. (eds.) E-Governance and Cross-boundary Collaboration: Innovations and Advancing Tools. IGI Global (2011).
2. Ojo, A., Estevez, E., Janowski, T., *Estrategic Planning for Electronic Governance – Process*, UNU-IIST Center for Electronic Governance, UNeGov.net School on Electronic Governance, Cucuta, Colombia, 4 al 6 de Septiembre de 2008.
3. Estevez, E., Janowski, T., *Government Information Sharing in Macao SAR*, UNU-IIST Center for Electronic Governance. (2010).
4. Estevez, E., Janowski, T., Marcovecchio, I., Ojo, A., *Establishing Government Chief Information Officer Systems - Readiness Assessment*. In: Bertot, J.C., Nahon, K., Chun, S.A., Luna-Reyes, L.F., and Atluri, V. (eds.) Proceedings of the 12th Annual International Conference on Digital Government Research (dg.o 2011), USA. pp. 292–301, (2011).
5. Reino Unido, *Ley de Protección de Datos*, https://www.gov.uk/data-protection/the-data-protection-act.
6. Estados Unidos, *Ley de Libertad de Información*, http://www.foia.gov/.
7. Unión Europea, *Framework de la Comunidad para las Firmas Electrónicas*, http://europa.eu/legislation_summaries/information_society/other_policies/l24118_en.htm.
8. Rodriguez Pottella, M., *Manual de Planificación Estratégica para Instituciones Universitarias*, FEDUPEL (1997).
9. Australia (Queensland Government, Department of Local Government, Sport and Recreation), *Stakeholder Engagement Policy*, http://www.docstoc.com/docs/34203478/ Stakeholder-Engagement-Policy--.
10. De Kluyver, C.A, *Pensamiento Estratégico: Una Perspectiva para Los Ejecutivos*. Pearson Educación (2001).
11. Dawes, S. (1996), *Interagency Information Sharing: Expected Benefits, Manageable Risks*, Journal of Policy Analysis and Management, Vol. 15, No. 3, pp. 377-394, JSTOR.
12. Pardo, T., Cresswell, A., Dawes, S., and Burke, B. (2004), *Modeling the Social & Technical Processes of Interorganizational Information Integration*, Proceedings of the 37th Hawaii International Conference on System Sciences, dg.o, vol. 62, ACM Portal.
13. Bryson, J. *Strategic Planning for Public and Nonprofit Organizations: A Guide to Strengthening and Sustaining Organizational Achievement*, John Wiley & Sons, Inc., 2011.
14. Bermejo, P.H.D.S, and Tonelli, A.O., *Planning and implementing IT governance in Brazilian public organizations*, 44th Hawaii International Conference on System Sciences, HICSS-44 2010; Koloa, Kauai, HI; United States; 4-7 Enero 2011, pp. 1-10.
15. De Almeida Teixeira Filho, J.G, and De Moura, H.P., *MMPE-SI/TI (Gov) - Model to assess the maturity level of the IS/IT strategic planning of Brazilian governmental organizations*, Proceedings de PICMET: Portland International Center for Management of Engineering and Technology, USA, 31 Julio-4 Agosto 2011.
16. McGee, R. Information Technology (IT) S*trategic Planning for Libraries*, Library Management, vol, 27, issue 6-7, 2006, pp. 470-485.
17. Landsbergen, D., and Wolken, G. (2001), *Realizing the Promise: Government Information Systems and the Fourth Generation of Information Technology*, Public Administration Review, Vol. 61, Issue 2, April 2001, Wiley, InterScience.

Ciber-adicciones: Estudio del Comportamiento Poblacional por Simulación

Montesano, L.[1], Pollo Cattaneo, F.[1], Garcia-Martinez, R.[2]

1. Programa de Maestría en Administración de Negocios. UTN-FRBA
2. Laboratorio de Investigación y Desarrollo en Arquitecturas Complejas
Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús.
lmontesano@arnapsis.com.ar, rgm1960@yahoo.com

Resumen. La actual revolución digitalizadora de Internet podría actuar como medio de trasporte de productos y servicios de consumo inagotable hasta generaciones jóvenes, permeables a tecnologías donde la exposición desmedida podría estar gestando, en silencio, un conjunto de personas ciber-adictas. Este trabajo busca formular herramientas que permitan encontrar perfiles comunes, sobre aquellos usuarios de Internet, que consuman productos y servicios digitalizados, de modo que si experimentan algún patrón de conducta compulsiva al hacerlo, puedan minimizarse las consecuencias humanas e individuales en base a la investigación de las causas.

Palabras Clave: productos digitales, modelos de consumo, ciber-adicciones.

1. Introducción

Actualmente entre el 25% y 30% de los habitantes del planeta disponen de acceso a la Internet. Es un espacio perfecto para las ideas que da entrada a millones de personas que suben y bajan contenidos intentando democratizar el conocimiento humano [Krotoski, 2010]. Se presentan dos mundos, no excluyentes sino complementarios: uno real de recursos que se pueden ver y tocar y otro virtual en el que los bienes y servicios adoptan la forma digital [Rayport y Sviokla, 1995].

En medios de soporte electrónico pueden alojarse; libros, imágenes, videos, música, programas informáticos y otros entretenimientos, que son solicitados, entregados y comerciados por vía de ciber-mercados [Kotler y Keller, 2006], apuntalando al comercio electrónico. El consumidor aumenta su capacidad de acceso, obtiene mayor información y compara características, reduciendo e incluso eliminando a los intermediarios [Amit y Zott, 2001].

Este beneficio se debe en parte a la transacción de bienes y servicios digitalizados, donde la infinitud de stock, devenida en ceros y unos, los predispone como nuevas posibilidades que amplían las innovaciones de comercio. La “Ley de los activos digitales” establece, que a diferencia del mundo físico, no se agotan con su consumo [López Sánchez y Sandulli, 2002].

Los emprendimientos que tratan con productos y servicios digitalizados comercializables por Internet constituyen un campo de investigación de creciente

interés en el área de administración de negocios de base tecnológica [Andrade, 2000; van Hooft y Stegwee, 2001; Menascé, 2002; Young y Johnston, 2003; Janita Muñoz, 2005; Barua et al., 2007; Howson, 2008; Nascarella, 2009, Young-Ei y Jung-Wan, 2010, Ruíz y Palací, 2011; Elaluf-Calderwood et al., 2011].

La característica de extraterritorialidad de la red y la falta de normativa específica para sus contenidos, sea por televisión digital interactiva o telefonía móvil debe reconsiderarse. La identificación del ámbito espacial es irrelevante para estas nuevas tecnologías [KPMG, 2000].

Algunos países de condiciones fiscales paradisíacas comenzaron a alojar productos y servicios digitalizados controversiales [Acta de Comercio Libre y Zona Procesada, 1994]. Mientras otros de mayor desarrollo iniciaron los primeros debates sobre el tema [Acta de Prohibición del Juego en Línea, 1997]. Las pujas legales trasnacionales sobre estas metodologías de distribución continúan.

2. Delimitación del Problema

Considerando que se trata de una modalidad comercial novedosa [Kotler y Keller, 2006], que acciona sobre empresas y consumidores influenciados por la actividad publicitaría [Krotoski, 2010], la situación económica, ambiental y social, la disposición de tecnología, los servicios de comunicaciones y datos [Elaluf-Calderwood et al., 2011] en el marco de una gestión gubernamental [Aspis et al., 2006] podría ser necesario entender las consecuencias del consumo masivo de productos y servicios digitalizados.

Todo aspecto de la vida parece ser remodelado por Internet, creando riqueza única: conocimiento. Es un espacio de perpetua innovación donde los contenidos digitalizados en productos y servicios son la materia prima y la vez; la creación [Krotoski, 2010]. Algunos estudios indican que aquellos usuarios de la red enfrentan potencialmente una serie de nuevos conflictos inherentes al comportamiento. Compulsividad y trastornos de la personalidad son devenidos del consumo exagerado de contenidos en Internet, experimentando un tipo de ciber-adicción [Llinares Pellicer y Lloret Boronat, 2008]. Siguiendo a los autores citados se reconocen cinco categorías de ciber-adicción. La primera refiera a la desmedida búsqueda de información de todo tipo, le sigue el exceso de contacto en entornos (o redes) sociales, la adicción a los juegos (de apuestas o no), compras compulsivas y finalmente ciber-sexo [Llinares Pellicer y Lloret Boronat, 2008]. Algunos individuos como los niños, adolescentes y adultos vulnerables se encuentran en mayor grado de exposición que otros [Araya Dujisin, 2005].

La generación “Y” comprende a las personas nacidas entre los años 1981 y 2000. Esta generación se distingue por una actitud desafiante y retadora. Lo cuestionan todo, no quieren leer y sus destrezas de escritura son pésimas. Educados y técnicamente hábiles en el uso de nuevas tecnologías, son independientes, multiculturales y disponen de mayor tolerancia a las diferencias entre personas. Abiertos a temas polémicos y a familias no tradicionales [Fonseca, 2003] son propensos al consumo en masa.

Para estimar un segmento poblacional de generación "Y" se puede considerar la evolución demográfica de La Argentina. Las personas nacidas entre los años 1981 y 2001 son 8.310.650 millones de habitantes. Aquellos con plena capacidad de derecho, mayores de edad (21 años para la Ley Civil de La Argentina) los aproxima a la cantidad de 4.666.048 millones de habitantes [INDEC, 2010].

La actual revolución digitalizadora de Internet [Krotoski, 2010] podría actuar como medio de trasporte de productos y servicios de consumo inagotable [López Sánchez y Sandulli, 2002] hasta generaciones jóvenes, permeables a tecnologías [Fonseca, 2003] donde la exposición desmedida podría estar gestando, en silencio, un conjunto de personas ciber-adictas [Llinares Pellicer y Lloret Boronat, 2008]. Debe considerarse que con poco más que diez años de comercio electrónico, algunos de los productos digitalizados como libros, música, películas y aplicaciones móviles; constituyen ejemplos de negocios de alto crecimiento en el planeta [Wasserman, 2010].

El problema que se presenta es encontrar perfiles comunes, sobre aquellos usuarios de Internet, que consuman productos y servicios digitalizados, de modo que si experimentan algún patrón de conducta compulsiva al hacerlo, puedan minimizarse las consecuencias humanas e individuales en base a la investigación de las causas.

3. Solución Propuesta

Los nuevos mercados de información o ciber mercados [Kotler y Keller, 2006] son lo suficientemente grandes y representativos, más precisos que otras técnicas para la extracción de información difusa, como las encuestas y sondeos de opinión [Asur y Huberman, 2010]. Los flujos de producción, modificación, intercambio y remixación de información responden a lógicas propias de entes colectivos [García y Gertrudix, 2011]. Los servicios abiertos están configurando un modelo revolucionario de intercambio y producción de información en la red. [Glez, 2006].

En los últimos años es común aprovechar la potencia informática para predecir resultados de comportamiento social [Asur y Huberman, 2010]. Apoyados en teorías matemáticas podrían resolverse dilemas de carácter social dentro del marco de la teoría de juegos [Glance y Huberman, 1994]. La información correctamente procesada podría aportar una forma de sabiduría colectiva para predecir resultados del mundo real [Asur y Huberman, 2010]. Puede utilizarse la potencia computacional para desarrollar un conjunto de datos para estudiar algo específico, constituyendo una fuente primaria, dado que la recopilación es propia [Ander-Egg, 2003].

Al estudiar datos estimados de sobre una posible realidad podría considerarse que se tiene una muestra simulada [Tarifa, 2001], obtenida a través de un proceso que incluye diseñar un modelo sobre un sistema real para llevar a cabo experiencias con él, a fin de aprender sobre su comportamiento [Shannon, 1988].

El método de simulación de Montecarlo es una técnica que combina conceptos estadísticos de muestreo aleatorio con la capacidad que tienen las computadoras para generar números pseudo-aleatorios y automatizar cálculos [Kalos y Whitlock, 1986; Peña Sánchez de Rivera, 2001]. Al agregar información acerca del comportamiento de una muestra representativa, podría disponerse saltos incrementales en la calidad de los datos y ser considerados confiables para obtener posibles conclusiones. La

evaluación de los resultados de una simulación podría dar entendimiento sobre el conjunto de consecuencias provocadas por un hecho o actuación afectada [Shannon, 1988].

La solución propuesta, al problema descrito en esta tesis, consiste en simular el consumo de bienes y servicios digitalizados por Internet, generando una muestra representativa de perfiles de usuario, apuntalado sus datos individuales con información de fuentes fidedignas sobre la sociedad contemporánea de La Argentina y otros típicos de usuarios de La Internet.

Para formular los aspectos relevantes del sistema de estudio se desarrollará un modelo matemático-informático que analice el modo de satisfacer las necesidades del presente, sin comprometer a las generaciones futuras para satisfacer las suyas, integrando los aspectos económicos, ecológicos y sociales, que son dinámicos e interactúan entre si, influenciándose el uno con los otros dos y, además, enlaza el corto plazo con el largo plazo [Gardetti, 2003b].

El apoyar el paradigma conceptual en criterios de sustentabilidad podría colaborar con un crecimiento humano justo, conectado, prudente y seguro [Gardetti, 2003a] donde la innovación y el cambio tecnológico posiblemente permitan alcanzar un desarrollo sustentable, colaborando con usuarios, empresas oferentes y estados fiscalizadores.

4. Experimentos

Los datos creados en el Simulador de Ciber-adicciones, por el método de Montecarlo-sociabilizado con información contemporánea de La Argentina [INDEC, 2001; INDEC, 2009; INDEC, 2011; INDEC, 2012; Samuelson et al., 2003] y otros comunes a los usuarios de Internet [Adigital, 2012; Muñoz-Ramos Mas, 2012] podrían ser interpretados con diferentes técnicas estadísticas u otras avanzadas sobre explotación de información.

En este trabajo se presentan los resultados e interpretaciones sobre el comportamiento simulado de perfiles de usuario de Internet en base a una adaptación de los criterios de referencia sobre desordenes mentales en ludopatía [APA, 1995]. Se divide el impacto en los ejes: económico, ambiental y social que podría producir en la masa de usuarios las ciber-adicciones a Internet (IAT), a los juegos de apuestas por Internet (DSM), a las compras compulsivas por Internet (CBMS) y las redes sociales por Internet (BFAS).

4.1. Análisis sobre el Eje Económico

En términos de impacto sobre el eje económico, como se muestra en la Figura 1 “Curva de valor de impacto por ahogo financiero y comisión de actos ilegales”, el ahogo financiero (AF) se presenta creciente para el conjunto de perfiles de usuario, pudiendo colaborar al efecto los medios de pago y las posibilidades crediticias brindadas por el sistema bancario. Esto podría sostener una hipótesis referida al aumento sobre el consumo, hasta el ahogo, debido a las posibilidades crediticias y el dinero virtual. La comisión de actos ilegales (CAI) se presenta, para la media de

perfiles de usuario, con bajo impacto. Sobre el final, la curva se eleva y podría indicar que los pocos perfiles de usuario que llegan hasta el extremo de cometer actos delictivos, afrontan un fuerte impacto económico por el consumo de bienes y servicios digitalizados por Internet.

4.2. Análisis sobre el Eje Ambiental

Al analizar el criterio de preocupación recurrente (PR) por el consumo de bienes y servicios digitalizados por Internet, según la Figura 2 “Curvas de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet” se podría argumentar que la curva sobre aspectos ambientales de compras compulsivas (CBMS) supera en términos de impacto a las demás y podría ser el principal aportante a la preocupación recurrente. Siguen las gráficas sobre el uso de Internet (IAT), conexión a las redes sociales (BFAS) y uso de juegos de apuestas (DSM) que podrían generar un pensamiento continuo sobre el consumo y alterar algunas esferas del perfil de usuario relacionadas con la ansiedad.

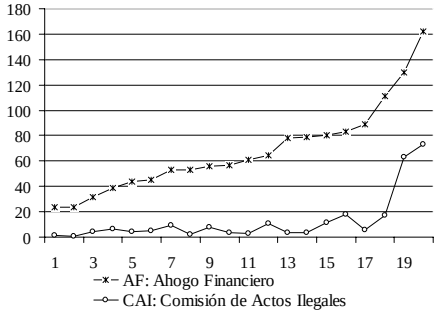


Fig. 1. Curva de valor de impacto por ahogo financiero y valor de impacto por comisión de actos ilegales. (Promedio de a 500 perfiles de usuario, valor de impacto)

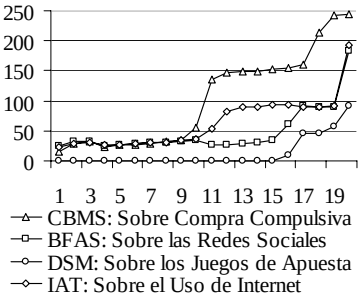


Fig. 2. Curvas de valor de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet (Promedio de a 500 perfiles de usuario, valor de impacto)

Las curvas de las variables independientes que se comparan en la Figura 3 “Curvas de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet” refieren al criterio de progresión del incremento (PI), pudiéndose argumentar que lo referido al uso de Internet (IAT) supera en términos de impacto y muestra un salto creciente y pronunciado para un grupo de perfiles de usuario. Esto podría indicar la necesidad de encontrarse conectado a Internet durante cada vez más tiempo para sentir confort. La progresión del incremento en las operaciones de compra (CBMS) crece en mayor medida que la participación de redes sociales (BFAS) y el uso de juegos de apuesta (DSM). Estas condiciones podrían relacionarse con la voluntad que presenta cada perfil de usuario para acotar estos consumos.

La Figura 4 “Curvas de impacto aportado por compras compulsivas, uso de redes sociales y uso de juegos de apuesta” analiza el comportamiento de las variables independientes sobre la intensión de retiro (IR) al consumir bienes y servicios digitalizados. Se podría argumentar que la curva para el uso de las redes sociales

(BFAS) presenta un pico de impacto por encima de las demás e indicaría un contenido difícil de abandonar para un algún grupo de perfiles de usuario. Para la media se presenta de bajo impacto. Al analizar la gráfica de los aportes sobre las operaciones de comercio en los portales de compra (CBMS) podría interpretarse alguna dificultad para retirarse y terminar la sesión, tal vez no tanto como con el uso de las redes sociales (BFAS). La curva sobre el uso de juegos de apuesta (DSM) se presenta de bajo impacto, pero un grupo de perfiles de usuario se muestra con dificultades para retirarse del contenido. Podría ser importante considerar si los excesos de tiempo para consumo de bienes y servicios digitalizados ocurren esporádicamente.

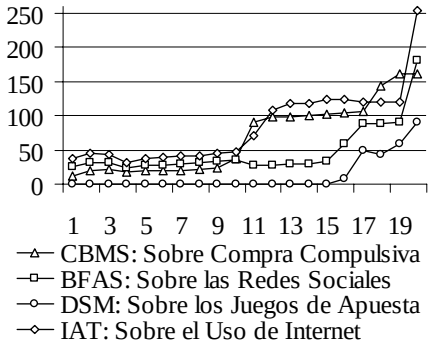


Fig. 3. Curvas de valor de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet (Promedio de a 500 perfiles de usuario, valor de impacto)

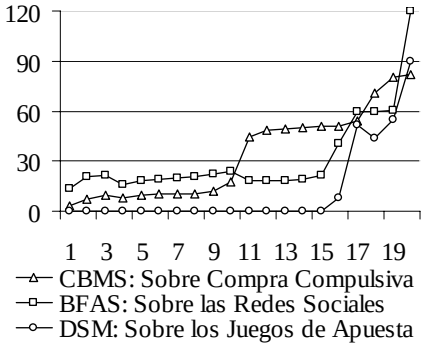


Fig. 4. Curvas de valor de impacto aportado por compras compulsivas, uso de redes sociales y uso de juegos de apuesta (Promedio de a 500 perfiles de usuario, valor de impacto)

Al analizar las curvas de las variables independientes según la Figura 5 “Curvas de impacto aportado por compras compulsivas, uso de juegos de apuesta y uso de Internet” respecto del criterio de tendencia a la repetición (TR) se podría argumentar que la curva sobre aspectos ambientales por compras (CBMS) se muestra por encima de las demás y podría indicar que la actividad comercial se repite fuertemente para un grupo de perfiles de usuario. Para la media se presenta de bajo impacto. Al analizar la gráfica de los aportes sobre el uso de Internet (IAT) podría interpretarse alguna dificultad para no reiterar la actividad, tal vez no tanto como con las compras (CBMS). La curva sobre el uso de juegos de apuesta (DSM) se presenta de bajo impacto, pero un grupo de perfiles de usuario se muestra con dificultades para controlar el uso reiterado de los juegos.

4.3. Análisis sobre el Eje Social

La Figura 6 “Curvas de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet” presenta a las curvas de las variables independientes sobre el criterio de perdida de control (PR) y podría argumentarse que la curva sobre aspectos sociales de compras compulsivas (CBMS) supera en términos de impacto a las demás siendo el principal aportante a la pérdida

de control. Siguen las gráficas sobre el uso de Internet (IAT), conexión a las redes sociales (BFAS) y uso de juegos de apuestas (DSM) que podrían generar un pensamiento continuo sobre el tema para olvidar temporalmente otros aspectos de la realidad.

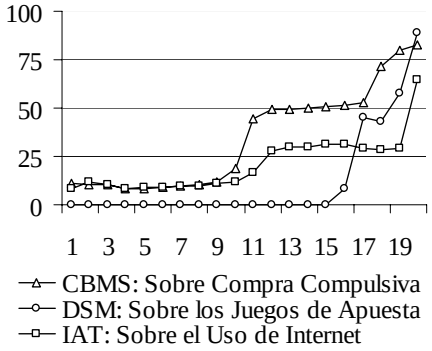


Fig. 5. Curvas de valor de impacto aportado por compras compulsivas, uso de juegos de apuesta y uso de Internet (Promedio de a 500 perfiles de usuario, valor de impacto)

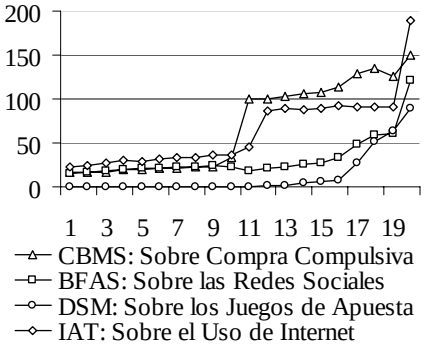


Fig. 6. Curvas de valor de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet (Promedio de a 500 perfiles de usuario, valor de impacto)

Al analizar las curvas de las variables independientes sobre el criterio de escape de la vida (EV) de la Figura 7 “Curvas de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet” se podría argumentar que la curva sobre aspectos sociales por el uso de Internet (IAT) supera en términos de impacto a las demás y podría ser el principal aportante al escape de la vida. Siguen las gráficas de conexión a las redes sociales (BFAS), compras compulsivas (CBMS) y uso de juegos de apuestas (DSM) que podrían presentar una serie de actividades que permitan al perfil de usuario olvidar sus obligaciones cotidianas y no afrontar los aspectos reales de la vida, pudiendo ser una actividad personal y secreta. La tendencia al ocultamiento (TO) sobre el consumo de bienes y servicios digitalizados se analiza en la Figura 8 “Curvas de impacto aportado por compras compulsivas, uso de juegos de apuesta y uso de Internet”. Las variables independientes podrían argumentar que la curva sobre aspectos sociales por el uso de Internet (IAT) se muestra por encima de las demás y es acompañada por la curva sobre compras (CBMS) pudiendo indicar que son actividades que tienden a ser ocultadas. La curva sobre el uso de juegos de apuesta (DSM) se presenta de bajo impacto, pero un grupo de perfiles de usuario se muestra con dificultades para controlar la actividad cayendo en el ocultamiento, incluso ante otros relacionados por sentimientos afectivos.

5. Conclusiones

Sobre los ejes económico, ambiental y social, se han presentado varios aspectos y criterios de estudio con dos posibles grupos de impacto: bajo y alto, lo que podría

sostener la hipótesis de que algunos perfiles de usuario no presentan atracción hacia el consumo masivo de bienes y servicios digitalizados por Internet o mantienen un buen nivel de control de consumo. Otros podrían presentar mayores dificultades.

La información obtenida en los experimentos podría ser procesada con otras técnicas de explotación de la información a fin de descubrir nuevas reglas e inferencias y complementar las gráficas e interpretaciones sobre las variables del modelo conceptual de impacto de consumo. Profundizar en esta linea de trabajo, permitirá obtener patrones de conductas de grupos que permitan a las empresas de Internet mitigar el posible impacto del consumo masivo de productos y servicios digitalizados en pos de mejorar la licencia para operar [Debeljuh, 2010] y desarrollar la perpetuidad de sus negocios.

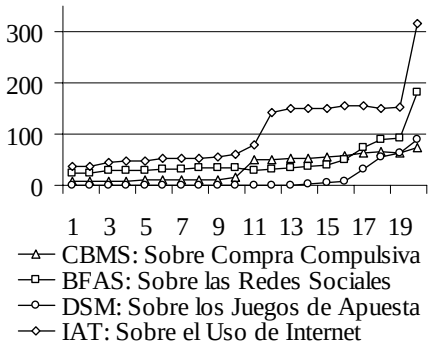


Fig. 7. Curvas de valor de impacto aportado por compras compulsivas, uso de redes sociales, uso de juegos de apuesta y uso de Internet (Promedio de a 500 perfiles de usuario, valor de impacto)

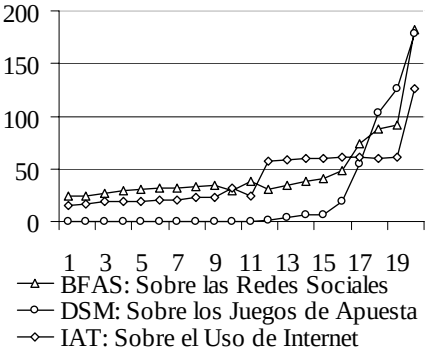


Fig. 8. Curvas de valor de impacto aportado por el uso de redes sociales, uso de juegos de apuesta y uso de Internet (Promedio de a 500 perfiles de usuario, valor de impacto)

6. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33B112 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

7. Referencias

Acta de Comercio Libre y Zona Procesada, 1994. “Free trade and processing zone act”, por el gobierno de Antigua Barbuda. http://www.antiguagaming.gov.ag/files/Antigua_and_Barbuda_Gaming_Regulations-Final.pdf (Última visita al sitio expuesto: 13 de julio de 2012).

Acta de Prohibición del Juego en Línea, 1997. “The Internet gambling prohibition act of 1997. 105th Cong. Hearing on H.R. 4777 Before the H. Comm. on the Judiciary and the Subcomm. on Crime, Terrpros, amd Homeland Security., 109th Cong. (2006) [hereinafter Ohr Statement] (statement of Bruce G. Ohr, Chief of Organized Crime and Racketeering Section, U.S. Dept. of Justice”.

http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:h4777ih.txt.pdf (Última visita al sitio expuesto: 13 de julio de 2012).
Adigital. 2012. Libro blanco de comercio electrónico. Guía práctica de comercio electrónico para PYMES. Asociación Española de la Economía Digital y Ministerio de Industria, Energía y Turismo de España. http://libroblanco.adigital.org//descarga.html (última visita al sitio expuesto: 15 de octubre de 2012)
Amit, R. y Zott, C. 2001. Value Creation in E-Business. Strategic Management Journal 22: 493–520. ISSN 1097-0266.
Ander-Egg, E. 2003. Métodos y técnicas de investigación: Técnica para recogida de datos e información social I. Grupo Editorial Lumen. ISBN 987-00-0301-X
Andrade, J. 2000. Formación de precios de los productos de información en redes digitales. Revista Venezolana de Gerencia. Universidad de Zulia. Venezuela. ISSN 1315-9984. http://revistas.luz.edu.ve/index.php/rvg/article/view/7881/7547 (Última visita 13/07/2012).
APA. 1995. Manual diagnóstico estadístico de los trastornos mentales. Asociación Americana de Psiquiatría. Versione española. ISBN: 84-458-0297-6. http://www.mdp.edu.ar/psicologia/cendoc/archivos/Dsm-IV.Castellano.1995.pdf (Última visita al sitio expuesto: 8 de julio de 2013).
Araya Dujisin, R. 2005. Internet, política y ciudadanía. 195: 56-71. ISSN 0251-3552.
Aspis, A., Pertusi, I., Nieva, H. 2006. Comercio Electrónico: e-commerce. Régimen contractual del comercio electrónico. Aspectos tributarios del comercio electrónico. Nuevas bases para gravar el e-commerce. Editorial Errepar. Bs. As. Argentina. ISBN-13 978-987-01-0570-1.
Asur, S. y Huberman, B. 2010. Predicting the future with social media. Social Computing Lab. HP Lab. http://arxiv.org/abs/1003.5699 (Última visita al sitio expuesto: 8 de julio de 2013).
Barua, A., Whinston, A., Konana, P. 2007. Assessing Internet Enabled Business Value: An Exploratory Investigation. Task Report. Center for Research in Electronic Commerce. Department of MSIS. McCombs School of Business. University of Texas at Austin. http://en.scientificcommons.org/43217363 (Última visita al sitio expuesto: 13 de julio de 2012).
Debeljuh, P. 2010. Ética empresarial en el núcleo de la estrategia corporativa. Editorial Cengage Learning. ISBN:978-987-1486-13-7.
Elaluf-Calderwood, S., Sorensen, C., Eaton B. 2011. Digital Innovation on Mobile Platforms: A Business Model Analysis. London School of Economics and Political Science. London. WC2A 2AE. UK. ACM 1-58113-000-0/00/0010. http://de2011.computing.dundee.ac.uk/ wp-content/uploads/2011/10/Digital-Innovation-on-Mobile-Platforms-A-Business-Model-Analysis.pdf (Última visita 13/07/2012).
Fonseca, J. 2003. Conociendo a la generación Y. Ponencia presentada en la 9na. Conferencia Anual del College Board. Universidad del Sagrado Corazón. Puerto Rico. http://www.collegeboard.com/ptorico/academia/diciembre03/conociendo.html (Última visita 13/07/2012)
García F. y Gertrudix, M. 2011, Naturaleza y características de los servicios y los contenidos digitales abiertos. Cuadernos de información y comunicación, ISSN: 1135-7991.
Gardetti, M. 2003a. Creando valor sustentable. Instituto de Estudios para la sustentabilidad corporativa. Buenos Aires. Argentina. Registro de propiedad intelectual Nro.: 274369. Copyright 2003.
Gardetti, M. 2003b. Desarrollo sustentable, sustentabilidad y sustentabilidad corporativa. Instituto de Estudios para la sustentabilidad corporativa. Buenos Aires. Argentina. Registro de propiedad intelectual Nro.: 274369. Copyright 2003.
Glance, N y Huberman, B. 1994. The dynamics of the social dilemmas. Scientific American. http://www.casos.cs.cmu.edu/education/phd/classpapers/Glance_Dynamics_1994.pdf (Última visita 8/07/2012)
Glez F., 2006. La Web 2.0: características, implicancias en el entorno educativo y algunas de sus herramientas", Departamento de Matemática, Universidad de Leon, España. http://www.iesevevirtual.edu.ar/virtualeduca/ponencias2006/La%20Web20_Santamaria.pdf (Última visita 10/03/2012)
Howson C. 2008. Business Intelligence. Estrategias para una implementación exitosa. Editorial McGraw-Hill Interamericana S.A. de C.V. Edición en Español. ISBN13 978-970-10-6759-8.
http://www.econ.unicen.edu.ar/attachments/1051_TecnicasIISimulacion.pdf (Última visita 8/07/2013)
INDEC. 2001. Censo Nacional de Población, hogares y viviendas. Definiciones de la base de datos.http://www.indec.gov.ar/redatam/CPV2001ARG/docs/Definiciones%20CD%20Base%20CNPHV2001_d.pdf (última visita al sitio expuesto: 15 de octubre de 2012)
INDEC, 2010. Evolución de la población argentina a través de los censos, Los censos de la población Argentina, Instituto Nacional de Estadísticas y Censo. http://www.censo 2010.indec.gov.ar/escuela.asp (Última visita al sitio expuesto: 13 de julio de 2012)
INDEC. 2009. Encuesta permanente de hogares. Diseño de registro y estructura para las bases de micro datos. Individual y Hogar. Instituto Nacional de Estadística y Censo.

http://www.santafe.gov.ar/index.php/web/content/download/80497/388465/file/EPH_disenoreg_09.pdf (última visita al sitio expuesto: 15 de octubre de 2012)
INDEC. 2011. Indec Informa. Instituto Nacional de Estadística y Censo. Año: 16. Nro.: 7. ISSN 0328-5804
INDEC. 2012. Censo nacional de población, hogares y viviendas 2010: censo del Bicentenario. Instituto Nacional de Estadística y Censo. Resultados definitivos. S^{e} B N^{o} 2. – 1ra ed. ISBN 978-950-896-421-2
Janita Muñoz, M. 2005. Los e-mercados, un nuevo modelo de mercado electrónico B2B. Departamento de economía aplicada y organización de empresas. Universidad de Extremadura. España. http://www.asepelt.org/ficheros/File/Anales/2005%20-%20Badajoz/comunicaciones/los%20e-mercados .pdf (Última visita al sitio expuesto: 13 de julio de 2012).
Kalos, M. y Whitlock P. 1986. Monte Carlo Methods. Vol I .Basics. John Wiley & Sons. New York.
Kotler, P. y Keller, K. 2006. Dirección de marketing. Duodécima Edición. Editorial Prentice Hall INC. ISBN 970-26-0763-9.
KPMG, 2000. The Economic Value and Public Perceptions of Gambling in the UK. Report for Business in Sport and Leisure. London: KPMG.
Krotoski, A. 2010. La revolución virtual. Open University on the BBC. The Open University. UK.http://www.open.edu/openlearn/science-maths-technology/engineering-and-technology/ technology/frontier-thinking/ou-on-the-bbc-the-virtual-revolution (Última visita 13/07/2012).
Llinares Pellicer, M. y Lloret Boronat, M. 2008. Ciber adicción: Los riesgos de Internet. Revista de Análisis Transaccional y Psicología Humanista, Nº 59. Vol: XXVI, Madrid. España. ISSN: 0212-9876
López Sánchez, J. y Sandulli, F. 2002. Evolución de los modelos de negocio en Internet: Situación actual en España de la economía digital. Universidad Complutense de Madrid. España. http://www.ucm.es/info/business/Documentos/articulos/030703.pdf (Última visita al sitio 13/07/2012).
Menascé, D. 2002. TPC-W: A Benchmark for E-Commerce. IEEE Internet Computing, 6(3): 83-87. ISSN 1089-7801.
Muñoz-Ramos Mas, M. 2012. Implicaciones socioeconómicas de las redes sociales en el mundo global. Tesina para Licenciatura en Publicidad y Relaciones Públicas. Facultad de Ciencias Sociales. Universitat Abat Oliba CEU.
Nascarella, M. 2009. Modelo de Agregación de Demandas Individuales con Reputación. Tesis de Magister en Ingeniería de Sistemas de Información. Escuela de Posgrado. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional. http://posgrado.frba.utn. edu.ar/investigacion/tesis/MIS-2009-Nascarella.pdf (Última visita al sitio expuesto: 13 de julio de 2012).
Peña Sánchez de Rivera, D. 2001. Deducción de distribuciones: el método de Montecarlo. Fundamentos de Estadística. Madrid: Alianza Editorial. ISBN: 84-206-8696-4.
Rayport, J. y Sviokla, J. 1995. Exploiting the virtual value chain. Harvard Business Review. http://hbr.org/product/exploiting-the-virtual-value-chain/an/95610-PDF-ENG (Última visita al sitio expuesto: 13 de julio de 2012)
Ruíz, M. y Palací, F. 2011. Variables cognitivas y psicología del consumidor. El modelo de la confirmación de expectativas en la actualidad. Boletín de Psicología, No. 103, 61-73. Departamento de Psicología Social y de las Organizaciones. Facultad de Psicología. Universidad Nacional de Educación a Distancia. http:// www.uv.es/seoane/boletin/ previos/N103-4.pdf (Última visita 3/04/2012).
Samuelson, P., Nordhaus W., Enrri D. 2003. Economía. McGraw –Hill, ISBN-13: 978-987-1112-02-9.
Shannon, R. 1988. Simulación de Sistemas. Diseño, desarrollo e implementación. Editorial Trillas. México. ISBN: 978-968-2426-73-5.
Tarifa, E. 2001. Teoría de modelos y simulación. Facultad de ingeniería. Universidad Nacional de Jujuy
van Hooft, F. y Stegwee, R. 2001. E-business strategy: how to benefit from a hype. Logistics Information Management, 14 (1/2): 44-53. ISSN 0957-6053.
Wasserman, A. 2010. Ciclo Conversando con los Líderes de los Negocios por Internet. Ponencia presentada en la conferencia nacional “E-Comerce Day”, 24 de septiembre de 2010. .Buenos Aires. Argentina. http://www.ecommerceday.org.ar/material/Wasserman.pdf (Última visita 27/04/2011).
Young, L. y Johnston, R. 2003. The role of the internet in business-to-business network transformations: a novel case and theoretical analysis. Information Systems and E-Business Management, 1(1): 73-91. ISSN 1617-9846.
Young-Ei, K. y Jung-Wan, L. 2010. Critical factors in promoting customer acceptance of and loyalty to online business management negree programs. African Journal of Business Management Vol. 5(1), pp. 203-211. Academic Journals. ISSN 1993-8233. http://www.academicjournals.org/ajbm/pdf/ pdf2011/ 4Jan/Kim%20and%20%20Lee.pdf (Última visita al sitio expuesto: 13 de julio de 2012).

Software e innovación: desarrollando productos con hardware y software flexible

Daniel Díaz, Sandra Oviedo, Leandro Muñoz, Francisco Ibañez,

Universidad Nacional de San Juan, Facultad de Ciencias Exactas Físicas y Naturales, Instituto de Informática y Departamento Informática

{ddiaz, soviedo, lmuñoz, fibannez}@iinfo.unsj.edu.ar

Resumen. Desde hace un tiempo la innovación se ha transformado en la fuente más importante de generación de valor y competitividad. En este contexto el software ha dejado de ser una tecnología de soporte oculta e invisible a los clientes para transformarse en el eje conductor del proceso creador de valor. Como consecuencia la ingeniería de software y la ingeniería de innovación tienen el desafío de integrarse y complementarse para adecuarse a los desafíos actuales. Este trabajo expone la importancia de la relación software e innovación y relata una experiencia académica que las vincula con el objeto de motivar el espíritu innovador de los alumnos, la misma se enfoca en el desarrollo de productos que tienen como base al software y hardware flexible.

Palabras claves: Software e Innovación, Software Flexible, Desarrollo de Productos Basados en Software

1 Introducción

Desde el surgimiento de la era de la computación, software, hardware e industria han vivido una sinergia. Cada cambio importante en el hardware y software ha sido influenciado por la industria y a su vez las nuevas tecnologías impulsadas por los cambios en el hardware y software han producido cambios en la industria.

Disponer de Software Flexible ha sido un anhelo desde el surgimiento del MRP (planificación de requerimientos de material) en los 1960 hasta la Cloud Manufacturing del 2010. Sin embargo el término software flexible ó software flexibilidad no tiene una definición concreta, esta depende de la perspectiva que se enfoque. Por ejemplo, desde una perspectiva sistémica Zhao [1] ha propuesto dos conceptos relacionados con la flexibilidad del software: la adaptabilidad del sistema y la versatilidad del sistema. Nelson y otros [2] definen a la flexibilidad de la tecnología como las características de la tecnología que permiten habilitar ajustes u otros cambios a los procesos de negocio. Desde la perspectiva del desarrollo de software la flexibilidad es una temática actual. Una gran cantidad de técnicas y métodos tiene a la flexibilidad como objetivo. Por ejemplo, el desarrollo dirigido por modelos, las líneas de productos de software, la programación generativa son paradigmas que intentan construir fábricas de software que permitan generar software a partir de solo describir un modelo que la fábrica de software interpreta para generar

una aplicación o software [3]. Esto es claramente un proceso de desarrollo flexible que permite modelar rápidamente los cambios que se producen en el ambiente y generar el código ejecutable que satisface a las nuevas necesidades.

Por otro lado, el Hardware flexible o hardware flexibilidad ha sido uno de los pilares que ha guiado la constante evolución del hardware. Desde el surgimiento de los circuitos integrados se ha buscado dispositivos electrónicos altamente customizables, adaptables a diversos usos, de interface sencilla, e interoperables. Todas estas características que definen un hardware flexible. Solo basta analizar la evolución de los microprocesadores para observar como estos han incrementado notoriamente su flexibilidad.

Hoy en día un nuevo movimiento denominado Hardware de Código Abierto ó Open Source Hardware está revolucionando la forma en la cual los productos serán diseñados, construidos y comercializados en un futuro. El Hardware de Código Abierto combina hardware y software flexible en una plataforma. Físicamente, este tipo de plataformas consiste en una placa basada en un microcontrolador que tiene entradas y salidas programables más un entorno de desarrollo de software. Es el software que permite transformar este hardware en un producto concreto. Este movimiento está facilitando el acceso a nuevas tecnologías de vanguardia a pequeños emprendedores quienes pueden desarrollar sus ideas tecnológicas con una muy baja inversión. Esto ha revalorizado aun más la creatividad y el poder innovador de una persona y está contribuyendo notablemente a establecer una nueva era, la era de la innovación.

Este artículo tiene por un objeto mostrar mediante el uso de una plataforma de hardware de código abierto cómo software e innovación es un mix interesante que puede ser utilizado para concretar ideas con alto contenido tecnológico utilizando un presupuesto de bajo costo.

2 Software e Innovación

Cada vez más la economía está pasando de una economía basada en el conocimiento a una economía basada en la creatividad y la innovación [4-6]. Un estudio realizado por Siemens pone de manifiesto que hoy en día hasta el 70% de los ingresos de una empresa es generado por productos o características que no existían hace cinco años [7]. Además de esto, el 90% de los gerentes de empresas en sectores como la aviación, del automóvil, el farmacéutico, y las telecomunicaciones consideran la innovación como algo esencial para alcanzar sus objetivos estratégicos [4]. "El dilema del innovador" [8] ya no es, en muchos casos, un dilema para las empresas que desarrollan productos, la innovación se ha convertido en una necesidad absoluta para hacer frente a los desafíos globales y las tendencias del futuro.

Una observación importante a realizar en este contexto es que el software es incrementalmente usado como el instrumento para hacer innovación: “el software es el motor de la innovación”. Esta tendencia no sólo es válida dentro de los nichos de mercado específicos, sino que el software es un elemento relevante en una amplia gama de sectores. Software ya no es una tecnología de apoyo, sino que este toma un papel esencial en el proceso de creación de valor.

Ser "el primero" es importante si se quiere ser innovador. Por eso, muchas empresas están compitiendo en una carrera por hacer punta en el mercado. El ciclo de vida del producto se está reduciendo a un ritmo constante, hoy en día, pocos son los productos con un ciclo de vida de un año o más.

2.1 Desarrollo Flexible y desarrollo de software flexible

La necesidad de ser el primero, la carrera por la punta del mercado, y el dilema de la innovación conducen a una necesidad de obtener flexibilidad. Desarrollo flexible es la capacidad de responder rápidamente a las nuevas necesidades del mercado y las peticiones del cliente. Se trata de aumentar la velocidad a la que las innovaciones y las ideas son llevadas al mercado. Las empresas sienten la necesidad creciente de ofrecer productos a tiempo. En este contexto y desde la perspectiva del desarrollo de software es necesario introducir el concepto y las técnicas necesarias para llevar a cabo el desarrollo flexible. Desde este punto de vista se puede definir al desarrollo de software flexible como el proceso que permite el desarrollo flexible.

2.2 Innovación y Software Flexible

Como se ha mencionado en algunos párrafos anteriores, la innovación se ha convertido en una necesidad absoluta para hacer frente a los desafíos globales y las tendencias del futuro. De esta necesidad surge el software flexible que es uno de los elementos necesario tanto para hacer frente a los cambios que producen la innovación como así también para ser innovadores.

3 Ingeniería de la Innovación e Ingeniería de Software

El principal desafío de la industria del software siempre ha sido entregar productos de software a tiempo, adecuados a presupuestos pre-establecidos y con una calidad aceptable, esta ha sido y es la tarea de la Ingeniería del Software. En este campo importantes logros se han alcanzado, mediante un conjunto de herramientas muy bien logradas en áreas tales como desarrollo de software, gestión de recursos tecnológicos, arquitectura de software, análisis de requerimientos, calidad de software, testing automático, entre otras.

A lo largo de la historia de la industria diversas estrategias y tácticas han sido planteadas para generar valor. Desde hace un tiempo se perfila la innovación como la más importante fuente de generación de valor y competitividad [9], es decir que para ser competitivas, las compañías deberán ser innovadoras. En un vasto sector industrial la innovación está dejando de ser una palabra para transformarse en una acción. Las empresas están llevando a la práctica la innovación mediante lo que se conoce como gestión de la innovación, esta encierra a un conjunto de prácticas tales como la generación de las ideas, gestión de las mejoras de productos, gestión del ciclo de vida de productos, etc. Desde la academia estas prácticas se han ordenado formando un proceso que se conoce como Ingeniería de la Innovación. Por ejemplo, en [10] se

describen ampliamente 16 prácticas que realizan las empresas mas innovantes, el autor las denomina prácticas de ingeniería de la innovación.

Un cambio importante en lo que respecta al software y la generación del valor está ocurriendo hoy en día y es que el software ha dejado de ser una tecnología de soporte oculta e invisible a los clientes para transformarse en el eje conductor del proceso creador de valor.

Como consecuencia la ingeniería de software y la ingeniería de la innovación tienen el desafío de integrarse y complementarse para adecuarse a los desafíos actuales. Desde la perspectiva de la ingeniería de la innovación existe la necesidad de conocer más acerca del software y sus procesos. Mientras que desde la ingeniería del software es necesario aprender más sobre el proceso innovador para aprovechar las nuevas oportunidades que está ofreciendo el software en la generación del valor a través de la creación de nuevos productos y servicios.

4 Plataformas de hardware abiertas

La OSHWA (Open Source Hardware Association) [11], o asociación de hardware de código abierto en su declaración de principios establece que “el hardware de código abierto es el hardware cuyo diseño se hace disponible públicamente para que cualquiera pueda estudiarlo, modificarlo, distribuirlo, hacer y vender el diseño o hardware basado en ese diseño. El código fuente del hardware, el diseño a partir del cual está hecho, está disponible en el formato preferido para realizar modificaciones en él ". Téngase en cuenta que el hardware de código abierto se refiere específicamente a compartir los archivos de diseño digital de objetos físicos; si bien se apoyan otras formas de compartir, creemos que es importante tener claro el significado de hardware de código abierto.

A pesar que el movimiento de código abierto es muy reciente hay una amplia comunidad de empresas, individuos y grupos que están diseñando y haciendo hardware de código abierto. El ejemplo más conocido es la plataforma de hardware y software flexible “Arduino”, que es objeto de estudio de este artículo. La mayoría de las impresoras 3D que están apareciendo en el mercado son open source hardware (RepRap, MakerBot, Tantilus). Recientemente, una plataforma de juegos denominada OUYA está por ser lanzanda al mercado este año [12]. En estos tres ejemplos presentados no solo el hardware es open source sino el software que permite que los dispositivos cobren vida también es open source. Una cuestión importante a tener en cuenta es que el open source es un fenómeno multiplicador de producto innovantes. Por ejemplo OUYA abrirá nuevas fronteras en el mercado de los videos juegos, de la televisión de alta definición vía internet, nuevos e innovadores productos surgirán a partir de OUYA. Este camino ya está demostrado con la plataforma Arduino. Hoy en día existe un numeroso conjunto de accesorios que se pueden adjuntar a un Arduino transformando a este en artefacto especializado. Así es posible transformar un Arduino en robot, en una alarma domiciliaria, en un control de riego. Estas plataformas son excelentes medios para explotar la creatividad y transformar ideas en producto innovantes.

5 Arduino

Arduino es un plataforma de hardware abierta basada en una tarjeta electrónica que posee entradas y salidas más un ambiente de desarrollo basado en el lenguaje Processing. Se puede usar para desarrollar objetos interactivos de funcionamiento independiente u objetos que se conectan a una computadora y que pueden interactuar con ella. Por ser un hardware open source la tarjeta puede ser construida por uno mismo o se puede comprar preensamblada. El ambiente de desarrollo o IDE (Integrated Development Environment) se puede descargar desde www.arduino.cc.

Existen diferentes modelos de Arduino, en Fig. 1 se puede observar el Arduino Diecimila. Se pueden ver las entradas y salidas, la conexión USB que permite conectarlo a la computadora para su programación, la conexión a fuente de alimentación y la descripción de la ubicación de cada uno del circuitos integrados.

El costo de un Arduino Duemilanove Atmega328 *"hecho en China"* es alrededor de los 25 dólares. En Argentina, el Arduino original *"made in Italy"* se puede conseguir por 200 pesos. El diseño de la placa de Arduino permite agregar "módulos" llamados "shields" concatenándolos, que expanden la conectividad y aplicaciones del sistema y/o reducen la carga computacional del microcontrolador. Los módulos más comunes son de GPS, tarjeta SD, ethernet, Xbee (wireless), bluetooth, touchshield, leds e I/O expandidos, entre otras.

El microcontrolador por defecto no posee sistema operativo (lo cual es lógico). Tan solo existe un bootloader que carga el programa y lo inicializa.

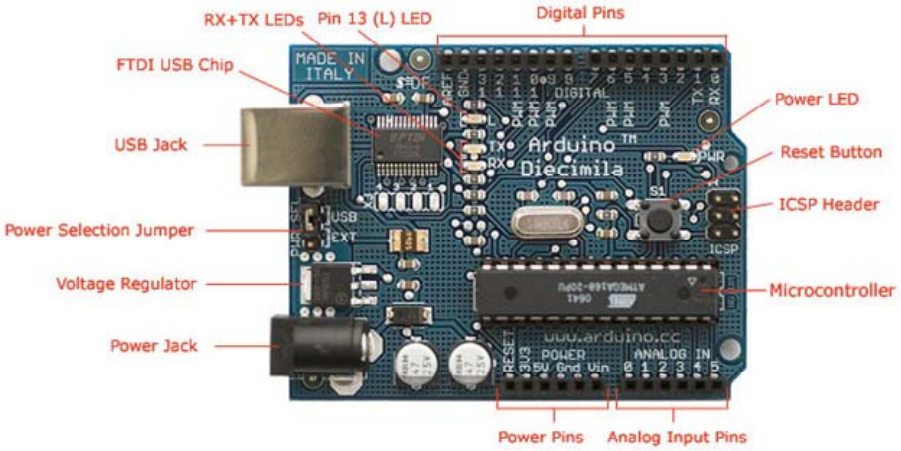


Fig. 1 Arduino diecimila. Imagen tomada de SparkFun.

En Fig. 2 se muestra el aspecto del ambiente de trabajo que permite programar la tarjeta Arduino.

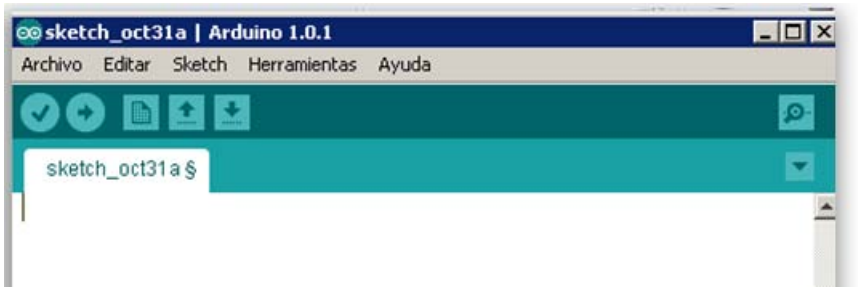


Fig. 2 Ambiente de desarrollo de software de Arduino

El proceso de generación de un producto con Arduino simplificado se inicia con una idea, la cual se plasmara en un prototipo. Se conecta el Arduino a la computadora que se utilizara para desarrollar el programa que transforma al Arduino en el producto deseado. Utilizando el IDE de Arduino compilamos el programa, si no existen errores se procede a desplegar el programa que está en la computadora al Arduino. A partir de este punto se puede comenzar con las pruebas correspondientes o testing. En Fig. 3 se muestra el proceso de construcción de un producto con Arduino.

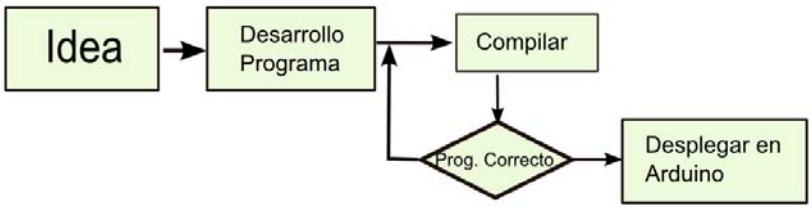


Fig. 3 Proceso de construcción de un producto con Arduino

En realidad el proceso de desarrollo es iterativo basado en prototipación, es decir, que se arriba al producto final mediante la prototipación incremental.

Arduino está siendo ampliamente utilizado para el desarrollo de productos a nivel prototipo. Es importante mencionar que Arduino favorece a la innovación, o a motivar el espíritu innovador, pero de ninguna manera el sólo hecho de usarlo asegura que se conseguirá un producto innovador.

6 Experiencias académicas. Innovación y software en las aulas.

En este apartado relatamos dos experiencias que relacionan software e innovación, que tuvieron por objetivo motivar el espíritu innovador utilizando la plataforma de hardware abierta Arduino.

La primera experiencia estuvo focalizada en el software con intención de ser utilizarlo en innovación, la segunda, se enfocó en temas de innovación utilizando el software como un camino para innovar. En oras palabras, en la primera experiencia

no se trabajo en la concepción de las ideas, éstas estaban preconcebidas, en la segunda experiencia se trabajo en la generación de ideas para prototipar con Arduino.
Se desarrollaron dos experiencias, en la modalidad de taller:

- Taller introducción a la programación de Arduino
- Taller de creatividad e innovación. incorporando hardware y software flexibles en la concepción de nuevos productos.

6.1 Taller introducción a la programación de Arduino

El objetivo del taller fue introducir a los alumnos en el conocimiento de la plataforma Arduino, diseño de prototipos básicos y programación con lenguaje Arduino para el diseño de sistemas y/o productos interactivos. El taller estuvo dirigido a alumnos de las carreras de informática e ingeniería.

Unos 20 estudiantes asistieron a este taller pertenecientes a las carreras de informática, ingeniería electrónica e ingeniería industrial. Para desarrollo de prácticas, se plantearon desafíos en los cuales se debía desarrollar tanto el componente de hardware como el componente de software, a fin de prototipar la solución. Los alumnos trabajaron en grupos de hasta 5 personas y tuvieron una semana para desarrollar su prototipo. A cada grupo se le proveyó de un kit de equipamiento básico y materiales para el desarrollo que constaba de: 1 tester, 1 plataforma Arduino, 1 experimentor, potenciómetro, juego de llaves, leds, resistencias, displays, etc.

Los desafíos fueron los siguientes:

Free Drink. A un innovador barman se le ocurrió la siguiente idea para incrementar las ventas en su bar: Presentar un escala visual de 5 luces de la más débil a la más intensa, conforme avanzan las ventas hay un algoritmo que enciende la siguiente luz en la escala (caso simple: cada $1000), así hasta llegar a la última. Cuando llega al color más intenso, se declara una vuelta de bebidas gratis y la escala vuelve al punto más débil.

Ecualizador de luces. En una obra de teatro para crear distintos ambientes se necesita que la iluminación cambie de amarillo a rojo, luego de rojo a azul y por último de azul a blanco de manera incremental y lenta. Es decir que en un determinado momento pueden existir dos colores con distinta intensidad.

Pedido de Peatonal. Se necesita programar dos semáforos para un paso peatonal en una ruta. Caso autopista: Este semáforo se activará solo cuando el peatón solicite la pasada apretando el botón ubicado en un pilar al costado de la ruta. Caso Urbano: Aquí cuando un peatón presiona el botón hay un algoritmo que determina cuándo darle luz verde al peatón, este tiempo va a depender del tráfico y del estado de los semáforos de la esquina.

Reloj de Leds. Con 6 leds diseñar un reloj de arena para un juego que marque 2 minutos. Podrías proponer otro modelo de reloj?

Temporizador programable. Dada una luz intensa, se desea que la luz se desvanezca en un cierto periodo de tiempo, este período debe ser regulado por el potenciómetro, tal como podría hacerlo un alumbrado exterior, cuando aclara la luz natural se apaga el alumbrado.

6.2 Taller de creatividad e innovación. Incorporando Hardware y Software Flexibles en la concepción de nuevos productos.

Se planteó como una continuación del taller anterior, ya que se trabajó considerando que los asistentes estaban iniciados en la tecnología de Arduino. Como el taller estaba abierto a todos los estudiantes se hizo una breve introducción a este tema, y se mostraron los prototipos resultantes del taller anterior a fin de poner a todos los asistentes en conocimiento.

El objetivo de este taller fue introducir conceptos de gestión de la innovación y técnicas de creatividad. Teniendo en cuenta que los participantes ya conocían la tecnología de Arduino, se propuso este taller para la generación de ideas a desarrollar con Arduino.

Con una convocatoria más amplia, hubo presencia de estudiantes de las carreras de informática, ingeniería electrónica, ingeniería industrial y diseño industrial. El taller se realizó con 18 estudiantes. Para el desarrollo de prácticas se propuso como eje temático el desarrollo de juguetes tecnológicos y se hicieron dos sesiones de creatividad aplicando dos de las técnicas presentadas.

6.3 Evaluación de las experiencias

Para los asistentes a los talleres, se pudo observar que las tecnologías de hardware abierto fueron una revelación, en el taller de Arduino, se lograron resultados sorprendentes para todos, se plantearon los desafíos y las soluciones fueron más allá de estos. Se establecieron vínculos y se valoró el trabajo multidisciplinario, todos fuimos testigos de cómo los unos enseñaban a los otros, cada uno explotando sus capacidades. Puede decirse que se formó una pequeña comunidad con vistas a seguir trabajando en las ideas de proyectos surgidas en las sesiones de creatividad.

Ambas experiencias fueron muy positivas, para docentes y alumnos. Desde el punto de vista de los docentes, entre otras cosas, permitieron conocer lo que podríamos llamar la apertura mental de los estudiantes, es decir, se pudo tener una noción de cuán abiertos mentalmente están los alumnos para proponer o aceptar nuevas ideas. Donde encontramos dificultades para previsualizar los productos en su completitud, solo se enfocaban en las áreas más pertinentes a su formación, los estudiantes informáticos ansiosos por resolver el tema relacionado al software y los estudiantes de ingeniería preocupados por la solución de hardware.

Esto se acentuó más en las sesiones de creatividad al punto que en algunos casos, para expresar una idea, en la instancia de pensamiento divergente, empezaban explicando la solución tecnológica para desarrollarla. Fue muy difícil lograr un pensamiento claramente divergente. Asimismo, se concibieron más de veinte ideas factibles técnicamente y muy novedosas.

También fue muy enriquecedora la experiencia para los docentes en otros aspectos, ya que fue la primera. Se ganó experiencia en dinámica de grupos, debiendo reconocer que fue muy difícil lograr la atmosfera adecuada en la sesión de creatividad, lo cual dejó la enseñanza que para esta actividad es mejor una localización que no sea en el mismo ámbito de la facultad, en un ambiente más propocio. En cuanto a lo institucional cuando se propuso la iniciativa, siendo totalmente extra curricular, hubo opiniones encontradas, apoyos y rechazos en los

colegas. En cuanto al vínculo logrado con los estudiantes, la experiencia fue muy buena, sirvió para ponernos a todos, docentes y estudiantes, en conocimiento del gran potencial que puede tener las personas cuando trabajan motivas, en conjunto, integradas con otras personas de otras especialidades.

7 Conclusiones

Se ha intentado poner de manifiesto la importancia de la relación software e innovación desde una perspectiva actual de la industria. La necesidad que existe que la ingeniería de software y la ingeniería de innovación se integren y complementen en pos de generar valor. Se ha expuesto en qué consiste una plataforma de hardware abierta y como esta se pueden transformar en un medio para explotar la creatividad y transformar ideas en producto innovantes. Enfocándonos en Arduino, una plataforma de hardware abierta se ha descripto una experiencia académica que vincula al software y a la innovación. El objetivo de la experiencia fue motivar el espíritu innovador de los alumnos enfocándonos en el desarrollo de productos que tienen como base al software y hardware flexible. Al evaluar las experiencias hemos denotado los aspectos más notables de las mismas.

8 Trabajos Futuros

Como trabajos futuros se propone el desarrollo de talleres que traten sobre herramientas de gestión de la innovación con el objetivo de seguir instalando la cultura de la innovación en la comunidad universitaria, el cual creemos que es el medio más propicio para la proliferación de nuevas ideas y generación de productos innovadores que se apoyan en el trabajo multidisciplinario.

Referencias

1 Zhao, J.L., Intelligent Agents for Flexible Workflow Systems. AIS Americas. Conference on Information Systems, Baltimore. Maryland, 1998.
2 Nelson, K.M., H.J. Nelson, and M. Ghods, Technology flexibility: conceptualization, validation, and measurement. HICSS97- Maui-Hawaii, 1997.
3 Greenfield, J. and K. Short, Software Factories: Assembling Applications with Patterns, Models, Framworks, and Tools.Wiley. 2004
4 Dehoff, K. and D. Neely, Innovation and product development: Clearing the new performance bar. , in Booz Allen Hamilton. 2004: http://www.boozallen.com/media/file/138077.pdf.
5 Nussbaum, B., R. Berner, and D. Brady, Get creative! How to build innovative companies., in Business Week. 2005: http://www.businessweek.com/magazine/content/05_31/b3945401.htm.
6 Leadbeater, C., We-think: Mass innovation, not mass production. 2008, London: Profile Books.
7 Rubner, J., Tuned in to today's megatrends, in Siemens's Pictures of the future. 2005.: http://w1.siemens.com/innovation/pool/en/publikationen/publications_pof/pof_fall_2005

/corporate_technology/interview_with_claus_weyrich/pof205_editorial_1326165.pdf. p. 90-91.
8 Christensen, C.M., The innovator's dilemma: When new technologies cause great firms to fail. . 1997, Boston: Harvard Business School Press.
9 Weil, T., Open innovation and the management of innovation. Global open innovation networks, OECD Mines Paristech, CERNA Innovation. , 2009 (http://www.oecd.org/sti/inno/42053837.pdf).
10 Boly, V., Ingènierie de l'innovation. Lavoisier - Hermes Science. Paris, France, 2008. ISBN 978-2-7462-1798-0.
11 Oshwa, Open Source Hardware Association. http://www.oshwa.org/, 2013.
12 OUYA, OUYA the console game. http://www.ouya.tv/, 2013.

Un Marco de Trabajo para la Integración de Arquitecturas de Software con Metodologías Ágiles de Desarrollo

Luis Vivas, Mauro Cambarieri, Nicolás García Martínez,
Marcelo Petroff, Horacio Muñoz Abbate.

Laboratorio de Informática Aplicada - Universidad Nacional de Río Negro
{lvivas, mcambarieri, ngarciam, mpetroff ,hmunoz}@unrn.edu.ar

Abstract. La construcción de software dentro de un marco metodológico ágil ofrece la posibilidad de contar con procesos livianos y simples, aplicando técnicas de programación que permiten expresar el concepto de agilidad, garantizando código de calidad desde el inicio del proceso de desarrollo. Algunas técnicas que han sido probadas en metodologías tradicionales de desarrollo de software son usualmente utilizadas en metodologías ágiles, como por ejemplo las pruebas de unidad. En este trabajo proponemos un marco de trabajo basado en una arquitectura en capas, permitiendo guiar el desarrollo de software por medio de una técnica de programación centrada en pruebas de unidad, la cual fue formalizada en una metodología ágil de desarrollo - eXtreme Programming. La contribución es un marco de trabajo que permite la integración de una arquitectura de software con la técnica de programación guiada por pruebas de unidad, y la identificación de tecnologías a utilizar para cada una de las capas de la arquitectura. El marco de trabajo propuesto se valida mediante un caso de estudio**.**

Keywords: Metodologías ágiles; eXtreme Programming (XP); Test Unitarios; Arquitectura de Software; Desarrollo Guiado por Pruebas

1 Introducción

Realizar un desarrollo de software con éxito y de calidad depende de varios factores como, por ejemplo, las personas seleccionadas, las herramientas y tecnologías a utilizar, la arquitectura de software y la metodología que guiará el proceso. Su correcta elección es un factor crítico de éxito.

La arquitectura de software brinda una visión abstracta de alto nivel, permitiendo plantear la reutilización y la evolución del código. Por otro lado, las metodologías ágiles permiten generar productos de calidad, basándose en la adaptabilidad del proceso de desarrollo de software para aumentar sus posibilidades de éxito por la flexibilidad y eficacia sobre las metodologías tradicionales.

Este trabajo explora como adaptar la arquitectura de software con la práctica del desarrollo guiado por pruebas (Test Driven Development - TDD) dentro de metodologías ágiles de desarrollo de software. En particular, considera la arquitectura de software en capas y la metodología de desarrollo ágil eXtreme Programming [1].

La contribución del mismo es mostrar la factibilidad del enfoque, presentando un marco de trabajo que incluye la selección de tecnologías que permiten su implementación. El marco propuesto se valida mediante un caso de estudio.

El trabajo está estructurado de la siguiente manera: La Sección 2 presenta los conceptos relacionados, incluyendo eXtreme Programming, arquitectura de software, desarrollo dirigido por pruebas (Test Driven Development - TDD) y pruebas unitarias. La Sección 3 explica el marco de trabajo propuesto y las tecnologías seleccionadas. A continuación, la Sección 4 valida el marco de trabajo propuesto a través de un caso de estudio y muestra como aplica TDD en el desarrollo de una de las capas de la arquitectura. La Sección 5 presenta otros aportes científicos en esta línea de investigación y discute la contribución de este trabajo. Por último, la Sección 6 brinda conclusiones y explica los trabajos futuros.

2 Conceptos Utilizados

Las siguientes secciones presentan los conceptos básicos utilizados en este trabajo, incluyendo: eXtreme Programming (Sección 2.1), arquitectura de software (Sección 2.2), test driven development (Sección 2.3) y pruebas unitarias (Sección 2.4).

2.1 eXtreme Programming

eXtreme Programming (XP) es una metodología ágil, descrita por Kent Beck [2], que centra sus prioridades en las personas y no en los procesos, alentando a los desarrolladores a responder a requerimientos cambiantes de los usuarios, aún en fases tardías del ciclo de vida del desarrollo. Se basa principalmente en la comunicación e interacción permanente con el usuario y en la programación de a pares (técnica de programación por parejas donde uno de los programadores escribe el código y el otro lo prueba, y luego se intercambian los roles). De esta forma, desde el principio, el código se prueba en base a requerimientos funcionales.

El proceso consiste de tres etapas: 1) *Interacción con el Cliente* - el cliente está disponible durante todo el proyecto para interactuar con el equipo de trabajo. De esta manera, se elimina la fase inicial de recolección de requerimientos, y éstos se van incorporando ordenadamente a lo largo del desarrollo; 2) *Planificación del Proyecto* – se basa en un diálogo continuo entre las partes involucradas en el proyecto, siendo el equipo el que estima el esfuerzo requerido para la implementación de cada funcionalidad; 3) *Diseño y Desarrollo de Pruebas* – el desarrollo guiado por pruebas es un enfoque ágil, donde para cada funcionalidad que se desea implementar, primero se escriben las pruebas y luego el código necesario para que la prueba sea exitosa. Una vez que el código cumple el test exitosamente, se amplía y continúa. De este modo, se realiza una integración continua, evitando un proceso más complejo al finalizar el proyecto [1]. El desarrollo guiado por pruebas formalizado en XP se conoce como Test Driven Development (TDD) [3].

2.2 Arquitectura de Software

La arquitectura de software conforma la columna vertebral de cualquier sistema y constituye uno de sus principales atributos de calidad [4]. El documento de IEEE Std 1471-2000 [5] define: "La Arquitectura de Software es la organización fundamental de un sistema encarnada en sus componentes, las relaciones entre ellos y el ambiente y los principios que orientan su diseño y evolución".

En particular, una arquitectura comúnmente usada es la definida en capas. En el caso de una aplicación empresarial puede dividirse en tres capas lógicas bien definidas [6]: 1) la capa de presentación, 2) la capa de negocio y 3) la capa de persistencia. El principio para la separación en capas es que cada una esconde su lógica al resto y solo brinda puntos de acceso a dicha lógica.

En la capa de presentación los objetos trabajan directamente con las interfaces de de negocios, implementando el patrón arquitectónico Model-View-Controller [6]. En este, el modelo (Model) es modificable por las funciones de negocio, siendo estas solicitadas por el usuario, mediante el uso de un conjunto de vistas (View) que solicitan dichas funciones de negocio a través de un controlador (Controller), que es quien recibe las peticiones de las vistas y las procesa.

La capa de negocio está formada por servicios implementados por objetos de negocio. Estos delegan gran parte de su lógica en los modelos del dominio que se intercambian entre todas las capas.

Finalmente, la capa de persistencia facilita el acceso a los datos y su almacenamiento en una base de datos.

2.3 Test Driven Development (TDD)

TDD es una técnica de programación que consiste en guiar el diseño de una aplicación, por medio de pruebas unitarias. Esta, cambia el orden tradicionalmente establecido, de manera que en primero se definen las pruebas y a partir de estas se va desarrollando la funcionalidad, repitiendo el ciclo, de acuerdo a lo que se espera que haga el software, por medio de integraciones y refactorizaciones – **(**una refactorización consiste en realizar modificaciones en el código con el objetivo de mejorar su estructura interna, sin alterar su comportamiento externo [7]) - continúas del desarrollo en los casos en que las pruebas no cumplan con el requerimiento.

Con esta técnica las pruebas constituyen la documentación del software que se está desarrollando.

2.4 Pruebas Unitarias

En los últimos años, los test unitarios han ido tomando cada vez más fuerza en el proceso de desarrollo de software y se integraron de una forma altamente productiva [8]. El desarrollo y ejecución de pruebas es una actividad fundamental en los proyectos de desarrollo de software, los cuales permiten mantener código de calidad durante el ciclo de vida del proyecto. Fundamentalmente, las pruebas unitarias

representan una alternativa para encontrar y corregir la mayoría de los errores de codificación [8].

Las pruebas unitarias como primer paso en el proceso de desarrollo son la base de la filosofía TDD, ya que resulta la mejor manera de producir código rápidamente y de calidad, permitiendo conducir el diseño, mediante la codificación de las citadas pruebas, antes de codificar interfaces o implementaciones [9].

3 Arquitectura de Software y Entornos de Trabajo (Framework)

Aplicando los conceptos explicados anteriormente, en esta sección presentamos un marco de trabajo que integra distintas tecnologías y frameworks disponibles en el mercado para la implementación de una arquitectura en capas, dirigida por pruebas unitarias en XP.

Para la implementación de la capa de presentación se propone el framework JSF [10], basado en el patrón MVC, el cual permite desarrollar rápidamente aplicaciones dinámicas creando páginas (vistas) y manejadores de vista (ManagedBean) de manera sencilla, simplificando el diseño de interfaces de usuarios. Una ventaja de esta elección es su capacidad de extensión para definir nuevos componentes e incorporar librerías existentes, como PrimeFaces[11] entre otras.

Para la implementación de la capa de negocio se propone la utilización de Spring Framework [12]. Un entorno de trabajo de código abierto, utilizado para la simplificación en el desarrollo de Aplicaciones Java Empresariales (JEE). Spring provee de un contenedor de objetos quien se encarga de administrar el ciclo de vida de los mismos, implementa los objetos de dominio como POJOs (*Plain Old Java Object* - sigla creada por Martin Fowler, Rebecca Parsons y Josh MacKenzie [13]) y representa objetos que son parametrizables a través de sus propiedades o constructores por medio de archivos de configuración u anotaciones. El contenedor maneja dos conceptos muy importantes para administrar las instancias de los objetos (POJOs): la Inversión de Control (IoC: Inversion of Control) y la Inyección de Dependencias (DI: Dependency Injection). El principio de Inversión de Control consiste en que el control de la construcción de los objetos no recae directamente en el desarrollador, sino que es otra clase o conjunto de clases las que se encargan de construir los objetos que se necesitan.[14].

Finalmente, para la capa de persistencia diseñada con el patrón DAO (Data Access Object) [15] se propone la utilización del Framework de código abierto Hibernate [16] como ORM (Mapeo Objeto Relacional). A través de Hibernate se realiza el mapeo del modelo de objetos (POJOS) a la base de datos relacional, mediante archivos declarativos o anotaciones en los objetos POJOS que permiten establecer estas relaciones. Hibernate está diseñado para ser flexible en cuanto al esquema de tablas utilizado, para poder adaptarse a su uso sobre una base de datos ya existente.

El marco de trabajo propuesto se ilustra en la Figura 1.

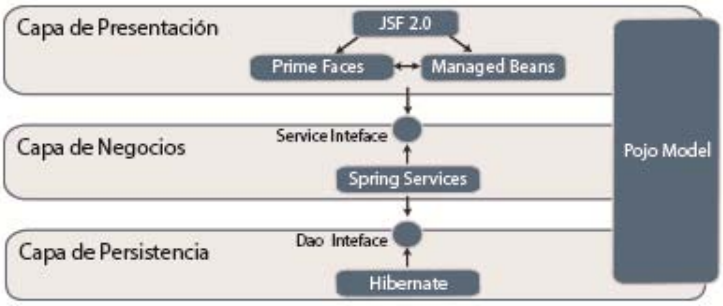


Figura 1. Arquitectura en Capas

4 Caso de Estudio

Una de las herramientas fundamentales para la realización de las pruebas de unidad más difundida en proyectos de software es conocida como JUnit [17]. Siendo una herramienta de prueba de código y el estándar para la técnica de TDD [9].

A continuación se presenta el caso de estudio mediante dos ejemplos, uno que se centra en depositar una suma de dinero en una cuenta bancaria y, el otro, que ejecuta una transferencia de un monto de dinero entre cuentas bancarias, guiados por TDD.

Para el primer ejemplo, la prueba de unidad permite probar la funcionalidad para el requerimiento "depositar dinero en la Cuenta". Una alternativa es la siguiente:

```
public void testDepositarEnCuenta() {
        // 1: Creamos un contexto de prueba
        Cuenta cuenta = new Cuenta();
        // 2: Ejecutamos el código bajo prueba.
        cuenta.depositar(100);
        // 3: Comprobamos el resultado.
        assertEquals(100, cuenta.getMonto(), 0.001);}
```

Prueba 1: Depositar Dinero en la Cuenta

Luego de la prueba de unidad se avanza en el diseño de la aplicación, por lo que se crea la clase "`Cuenta`" con su constructor y se implementa el método "`depositar`", el cual recibe un valor dado, para ello se modifica la estructura de la clase (el diseño) agregando el atributo "`monto`" que representa el valor.

El siguiente caso de prueba de unidad, permite comprobar la funcionalidad para el requerimiento "transferencia de dinero entre cuentas".

```
public void testTransferenciaEntreCuentas() {
   Cuenta cuenta1 = new Cuenta(100d);
   Cuenta cuenta2 = new Cuenta(100d);
   TransferenciaService transfService = new TransferenciaService();
   transfService.transferir(cuenta1, cuenta2, monto);}
```

Prueba 2: Transferencia de Dinero entre Cuentas

En este caso de prueba, se identifica un nuevo objeto de negocio, "TransferenciaService" que transfiere monto entre cuentas. Este implementa el método "transferir". La operación incluye a un nuevo objeto de negocio "CuentaService" que permite la actualización de las cuentas.

```
public void transferir(Cuenta cuenta1, Cuenta cuenta2, Double monto) {
  cuenta1.extraer(monto);
  cuenta2.depositar(monto);
  cuentaService.guardar(cuenta1);
  cuentaService.guardar(cuenta2);}
```

Implementación 1: Método Transferir de TransferenciaService

Siguiendo con la definición de la arquitectura propuesta, los objetos diseñados "CuentaService" y "TransferenciaService" se implementan en la capa de Negocio. De acuerdo al método "transferir" surge la necesidad de persistir los objetos "Cuenta", para ello se define el siguiente caso de prueba:

```
public void testGuardarCuenta() {
          CuentaService cuentaService = new CuentaService();
          Cuenta cuenta = new Cuenta();
          cuenta.setMonto(100);
          cuentaService.guardar(cuenta);}
```

Prueba 3: Guardar los Cambios de Cuenta

De la implementación del método `cuentaService.guardar(cuenta)`, citado, surge la necesidad de diseñar un nuevo objeto, llamado "CuentaDao", que se encarga de almacenar la "cuenta" en una fuente de datos. Este nuevo objeto se implementa en la capa de persistencia de la arquitectura.

```
public void guardar(Cuenta cuenta) {
        cuentaDao.guardar(cuenta);}
```

Implementación 2: Método Guardar de CuentaService

De los casos de prueba anteriores se realiza el diseño, de acuerdo al proceso de la técnica TDD, donde los requerimientos se traducen en pruebas. Es necesario, en ciertas oportunidades, que los objetos diseñados se comuniquen con otros para cumplir con su objetivo (función), para lo cual es importante simular la interacción entre ellos, sin llegar a construir el objeto real, de esta manera se logra que las mismas se realicen de forma independiente en cada una de las capas.

Las secciones a continuación explican, en detalle, la herramienta de pruebas para simular los objetos en la capa de negocio y las diferentes estrategias para aplicar pruebas unitarias en cada capa de la arquitectura.

4.1 Capa de Presentación

Esta capa contiene los manejadores ("ManagedBean"), que interactúan con otros objetos para colaborar con las acciones llevadas a cabo en la vista. La comunicación que existe con la capa subyacente es a través de la implementación de los objetos

Mocks (falsos) de la capa de Negocio utilizando la herramienta JMock. La Figura 2 muestra la implementación de las pruebas de unidad con JMock en la capa de presentación (`ManagedBean`).

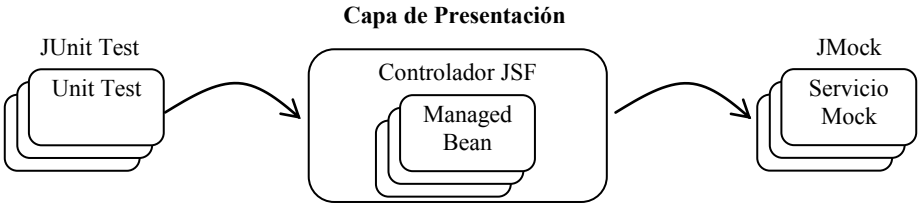


Figura 2. Implementación de las Pruebas en la Capa de Presentación

4.2 Capa de Negocio

A fin de realizar los test en la capa de negocios, se plantea la técnica denominada Mock Test. La utilización de esta técnica permite que las pruebas sean unitarias en lugar de que sean pruebas de integración (utiliza todos los componentes reales de los que depende). El desarrollo de pruebas unitarias en esta capa propone aislarla de los objetos de acceso a datos (DAO) y simular la implementación de los mismos con objetos Mocks (falsos).

Existen múltiples herramientas que permiten la creación de Objetos Mocks, como por ejemplo: MockObjects [18], jMock:[19], Mockito [20], EasyMock [21]. El diseño guiado en esta capa se realiza con la herramienta jMock, la guía de pasos para utilizar esta herramienta incluye: 1) declarar un contexto para la prueba, 2) crear los mocks dentro del contexto, 3) crear las expectativas (el comportamiento que se espera de los mocks), 4) ejecutar el código bajo prueba, y 5) comprobar si se han cumplido todas las expectativas.

A continuación, se describe el caso de prueba utilizando la herramienta JMock, siguiendo los pasos arriba descriptos:

```
//0. Crear un contexto.
Mockery context = new Mockery();

//1. Crear los mocks dentro del contexto a partir de las interfaces.
CuentaService cuentaService = context.mock(CuentaService.class);
TransferenciaService transfService = context.mock(TransferService.class);
public void testTransferirMontoEntreCuentas() {
 final Cuenta cuenta1 = new Cuenta(100);
 final Cuenta cuenta2 = new Cuenta(100);

//2. Definir las llamadas que esperan los mocks y los valores devueltos.
 context.checking(new Expectations() {
     {       oneOf(cuentaService).guardar(cuenta1);
             oneOf(cuentaService).guardar(cuenta2);
     }
  });

//3. Crear el objeto de la clase e invocar el método bajo prueba.
     transfService.transferir(cuenta1,cuenta2,50);
```

```
//4. Verificar que el comportamiento indicado en context se haya cumplido
 context.assertIsSatisfied();}
```

Prueba 4: Descripción del Caso de Prueba en JMock

Seguidamente se asegura la persistencia de la transferencia que se realiza utilizando objetos Mocks en el objeto `CuentaService`.

```
public void testGuardarCuenta() {
  final Cuenta cuenta = new Cuenta();
  final CuentaDao dao = context.mock(CuentaDao.class);
  context.checking(new Expectations() {{oneOf(dao).guardar(cuenta);} });
  cuentaService.setDao(dao);
  mockery.assertIsSatisfied();}
```

Prueba 5: Guardar en Cuenta utilizando JMock

La Figura 3 muestra como se implementan las pruebas de unidad con JMock en la Capa de Negocios.

Figura 3. Implementación de Pruebas de Unidad en la Capa de Negocios

4.3 Capa de Persistencia

La implementación de la capa lógica de persistencia, DAO (Data Access Object), se guía por pruebas, interactuando con un elemento externo (Base de datos). La utilización de DbUnit, como una extensión de JUnit, permite interactuar con un conjunto de datos de prueba y, también, dejar la base de datos en un estado conocido antes y después de cada ciclo de prueba, esto con el fin de prevenir que datos corruptos queden en la base de datos ocasionando problemas a los ciclos siguientes. Básicamente DbUnit usa archivos XML para cargar datos en la base de datos (dataset). [9]. La Figura 4 ilustra cómo se implementan las pruebas de unidad con DBUnit en la capa de persistencia (Dao).

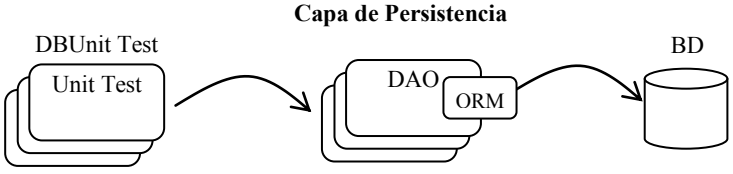


Figura 4. Implementación de Pruebas de Unidad en la Capa de Persistencia

5 Trabajos Relacionados

Este trabajo está basado en tareas de investigación que analizaron los aportes científicos que se encuentran en esta misma línea. En particular, se investigaron trabajos relacionados con la integración de las metodologías ágiles con arquitecturas de software que pudieran brindar un marco de calidad con un enfoque riguroso que resultara en mejoras de los procesos de diseño y de desarrollo de aplicaciones. Algunos de los trabajos más importantes se discuten a continuación.

Urquiza Yllescas, et al [22] plantea actividades de desarrollo que integran tácticas y estrategias de diseño de la arquitectura de software en metodologías agiles, desde una perspectiva general. Breivold y sus colegas [23] realizan una meta-investigación sobre publicaciones científicas que relacionan el desarrollo ágil y la arquitectura de software, concluyendo en la falta de evidencia científica para muchas afirmaciones sobre agilidad y arquitectura, resaltando la necesidad de estudios empíricos para demostrar ventajas y desventajas de aplicar un método ágil. En [24], se analizan dos casos de estudios adoptando patrones de diseños arquitectónicos en el desarrollo ágil de software para aplicaciones móviles. Resaltando la utilidad de una arquitectura basada en patrones y soluciones probadas. Finalmente, [25] presenta un arquitectura dirigida por modelos (MDA) y el proceso de XP, analizando los argumentos a favor y en contra, proponiendo una nueva arquitectura que supere las limitaciones identificadas.

Comparando el estado del arte con los resultados de este trabajo, nuestra tarea consistió en desarrollar, con el enfoque de TDD, una nueva perspectiva de aplicación ágil relacionándola con una determinada arquitectura e identificando un conjunto de tecnologías para su implementación, aportando, a nuestro criterio, evidencia práctica para su aplicación. De este modo, aportamos contribución empírica para demostrar las ventajas de combinar metodologías agiles con arquitecturas de software, como lo requerían las conclusiones de trabajos relacionados [22].

6 Conclusiones y Trabajos Futuros

Este trabajo presentó un marco de trabajo para el desarrollo de aplicaciones basado en una arquitectura en capas, aplicando la técnica de TDD para el diseño y desarrollo de cada una de las capas. La ventaja radica en producir capas de software altamente cohesivas, donde un nuevo requerimiento no impacta en el comportamiento entre cada una de ellas. La técnica aplicada sobre la arquitectura implica un cambio de mentalidad en el desarrollo de software, lo que permite implementar el código necesario para resolver cada caso de prueba concreto. Este enfoque permite reducir los típicos bloques de código que usualmente se agregan “por las dudas”, característica que habitualmente ocurre con metodologías de desarrollo tradicional.

Trabajos a futuro incluyen extender el marco de trabajo aplicando otras técnicas de desarrollo ágil como BDD (Behaviour Driven Development) y ATDD (Acceptance Test Driven Development) que permitirán diferentes posibilidades de integración dentro de nuestro ciclo de pruebas.

Referencias

[1] Mendes Calo, K, Estevez, E. and Fillottrani, P. "*Un Framework para Evaluación de Metodologías Agiles*" http://sedici.unlp.edu.ar/handle/10915/21086.

[2] Beck, K. "*Extreme Programming Explained. Embrace Change",* (1999).

[3] Wells, D., "*Extreme Programming Unit Tests*", disponible en: http://www.extreme programming.org/rules/unittests.html (accedido 01/06/2013).

[4] Clements, P., et al, "*Software Architecture in Practice*", Pearson Education, (2003).

[5] IEEE Standards Association, "*1471-2000 - IEEE Recommended Practice for Architectural Description for Software-Intensive Systems*", available at: http://standards.ieee.org/findstds/standard/1471-2000.html.

[6] Fowler, M. "*Patterns of Enterprise Application Architecture*", Addison-Wesley, (2002).

[7] Fowler, M., "*Refactoring: Improving the Design of Existing Code*", Addison-Wesley Longman, Inc., (1999).

[8] Johnson, R. and Hoeller, J., "*Expert One-on-One J2EE Development without EJB*", Wiley Publishing, (2004).

[9] Sam-Bodden, B., "*Beginning Pojos – From Novice to Professional*", APress, (2006).

[10] Oracle, "*Java Server Faces*", disponible en: http://java.sun.com/j2ee/javaserverfaces/ (accedido 02/07/2013).

[11] Prime Faces, "*PrimeFaces Ultimate JSF Component Suite*", disponible en: http://primefaces.org (accedido 13/03/2013).

[12] GoPIvotal, Inc., "*Spring Framework*", http://www.springframework.org/ (accedido 21/05/2013).

[13] Fowler, M., "*Plain Old Java Object (POJO)*", disponible en: http://www.martin fowler.com/bliki/POJO.html (accedido 07/06/2013).

[14] Johnson, R., et al, "*Professional Java Development with the Spring Framework*", Wiley Publishing Inc, (2005).

[15] Oracle, "*Data Access Object (DAO)*", disponible en: http://java.sun.com/blueprints/ corej2eepatterns/Patterns/DataAccessObject.html (accedido 21/03/2013).

[16] JBoss Community, "*Hibernate*", http://www.hibernate.org/ (accedido 22/03/2013).

[17] GitHub, "*JUnit*", www.junit.org/ (accedido 06/05/2013).

[18] Mock Blog, "*Mock Objects*", http://www.mockobjects.com (accedido 02/06/2013).

[19] GitHub, "*JMock*", http://www.jmock.org (accedido 16/06/2013).

[20] "*Mockito*", https://code.google.com/p/mockito/ (accedido 02/05/2013).

[21] Freese, T. and Tremblay, H. "*Easy Mock*", http://www.easymock.org (accedido 28/05/2013).

[22] Urquiza Yllescas, J.F., et al, "*Las Metodologías Agiles y las Arquitecturas de Software*". Coloquio Nacional de Investigación en Ingeniería de Software y Vinculación Academia-Industria 2010, 29-Setiembre al 1-Octubre 2010, León, Guanajuato, Mexico.

[23] Breivold,H.P., Sundmark, D., Wallin, P. and Larsson, S., "*What Does Research Say about Agile and Architecture?"* , en Proceedings of the 2010 Fifth International Conference on Software Engineering Advances (ICSEA), USA, (2010).

[24] Ihme, T. and Abrahamsson, P., "*The Use of Architectural Patterns in the Agile Software Development of Mobile Applications*", en Proceedings of International Conference on Agility. Helsinki, Finland, (2005).

[25] Guha, P., et al, "*Incorporating Agile with MDA Case Study: Online Polling System*", International Journal of Software Engineering & Applications (IJSEA), Vol.2, No.4, pp. 83-96, (Oct. 2011).

Herramienta de gestión de trazabilidad de requerimientos en proyectos de software

Alfredo Villafañe[1], María de los A. Ferraro[1], Yanina Medina[1], Cristina Greiner[1], Gladys Dapozo[1], Marcelo Estayno[2]
[1]Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste. Corrientes. Argentina
afv0185@hotmail.com, mafferraro@hotmail.com, {yanina, cgreiner, gndapozo}@exa.unne.edu.ar
[2]Departamento de Informática. Facultad de Ingeniería. Universidad Nacional de Lomas de Zamora. Buenos Aires, Argentina
mestayno@gmail.com

Abstract. La trazabilidad en la Ingeniería de Software es una práctica de control que contribuye a la construcción de un producto en el dominio de la solución lo más fiable y ajustado a las necesidades expresadas por el cliente. Se presenta una herramienta que permite administrar proyectos de software y realizar el seguimiento de los requerimientos en cada una de las fases de desarrollo, destacando la jerarquía que presentan sus relaciones. Además permite obtener documentación del proyecto que cumple con el estándar IEEE 830 e integra la metodología NDT. Como fortalezas de la herramienta se destacan la capacidad de seguir el ciclo de vida de un requerimiento, desde el origen de la especificación hasta la prueba del mismo, la capacidad de descubrir dependencias y conflictos entre requerimientos, y la capacidad de mejorar la comprensión del sistema en su totalidad, características que contribuyen a facilitar el desarrollo y mantenimiento del mismo.

Keywords: Ingeniería de Requerimientos. Gestión de proyectos de software. Trazabilidad de requerimientos.

1 Introducción

La Ingeniería de Requerimientos cumple un papel primordial en el proceso de desarrollo de software, ya que se centra en la definición del comportamiento del sistema. Su objetivo es la definición clara, consistente y compacta de las especificaciones correctas que definen el comportamiento del sistema con el fin de minimizar los problemas que se presentan en el desarrollo de software y que afectan a la calidad del producto final. La captura correcta de los requerimientos contribuye a la calidad del software dado que permite definir con precisión las condiciones que éste debe cumplir.

La trazabilidad en la Ingeniería de Software es una práctica de control que ayuda a obtener el producto en el dominio de la solución lo más exacto y fiable posible a las necesidades expresadas por el cliente en el dominio del problema. La trazabilidad está condicionada por los cambios y las validaciones que los participantes del proyecto hagan al sistema durante el proceso de desarrollo [1]. Según el estándar IEEE 830-1998, la trazabilidad es la habilidad para seguir la vida de un requerimiento en ambos

sentidos, hacia sus orígenes o hacia su implementación a través de las especificaciones generadas durante el proceso de desarrollo. Es un factor de calidad.

En el desarrollo de aplicaciones web, el requerimiento está inmerso en un proceso de ingeniería más amplio y detallado. La existencia de una importante estructura de navegación obliga a un desarrollo preciso de este aspecto que garantice que el usuario no se "pierda en el espacio navegacional del sistema" [2]. Estas características particulares requieren atención en la fase de especificación de requerimientos [3].

NDT (*Navigational Development Techniques*) [4][5] es una técnica para especificar, analizar y diseñar el aspecto de la navegación en aplicaciones web. Comienza con la captura de requerimientos y estudio del entorno, luego se definen los objetivos del sistema y en base a ellos se definen los requerimientos que el sistema debe cumplir para cubrir los objetivos marcados. Finalmente se desarrolla una matriz de trazabilidad que permite evaluar si todos los objetivos han sido cubiertos en la especificación.

Debido a la heterogeneidad de los usuarios de una aplicación web, es necesario considerar las necesidades de los diferentes actores implicados en la misma para determinar las características que la aplicación debe cumplir para satisfacerlas [6]. Aunque en la actualidad existen propuestas para la especificación de requerimientos web [7][8], la mayoría sólo proponen un conjunto de guías de diseño informales para la derivación manual de modelos conceptuales a partir de los requerimientos [9].

La trazabilidad es clave para conseguir una exitosa gestión de requerimientos, sin embargo no hay un consenso respecto de las prácticas con el que el proceso de trazabilidad debe llevarse a cabo [10].No existen estándares asociados al proceso de trazabilidad que ayuden a determinar qué tipos de artefactos y de enlaces considerar. Se considera como una medida de la calidad del software y es tenida en cuenta en modelos como CMMI (Capability Maturity Model Integration) [11].

De acuerdo al INTECO (Instituto Nacional de Tecnologías de la Comunicación) en su guía de Gestión de Requisitos, un requerimiento es trazable si se pueden identificar todas las partes del producto existente relacionadas con ese requisito [12].
Todos los requisitos deberían ser trazables para mantener consistencia entre los distintos documentos de un proyecto.
Es importante conocer aspectos de los requisitos tales como:

- Su origen (Quién los propuso)
- Necesidad (Por qué existe)
- Relación con otros requisitos (Dependencias)
- Relación con otros elementos (Dependencias)

Para aportar a una mayor sistematización en el desarrollo web, este grupo de investigación elaboró una propuesta metodológica para la especificación de requerimientos de aplicaciones web, basada principalmente en una plantilla que considera lo estipulado por el estándar IEEE 830-1998 e incluye las características particulares de los requerimientos web basados en NDT, y elementos trazables para facilitar el rastreo de los requerimientos y el impacto de los cambios [13].

Como continuación del trabajo anterior, se propuso elaborar una aplicación que implemente la propuesta metodológica y permita facilitar la evaluación del impacto de los cambios introducidos, minimizando las incoherencias producidas por el incorrecto control de cambios, y favorecer la comunicación entre los stakeholders en el desarrollo y mantenimiento del software.

2 Metodología

En [13] se propuso una plantilla para especificación de requerimientos, cuyos componentes pueden observarse en la figura 1. Presenta características del estándar IEEE 830 combinadas con la metodología NDT, reforzando además la trazabilidad de los componentes críticos, a través del uso de matrices. Cuenta con información concreta y precisa para asegurar calidad desde la trazabilidad, sin perder de vista componentes hoy casi imprescindibles en el desarrollo web, como los provistos por NDT.

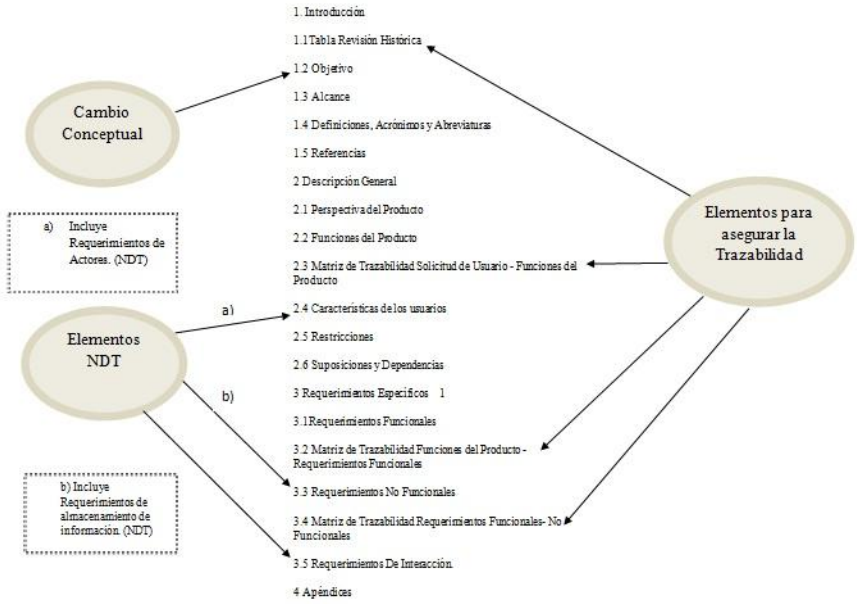


Fig. 1: Plantilla reformulada del estándar IEEE 830

Como soporte tecnológico de esta plantilla se diseñó una aplicación cuyo modelo se especifica en la figura 2 mediante el diagrama de casos de uso de la aplicación.
Se consideró la arquitectura **cliente servidor,** dado que las tecnologías de hardware, software, base de datos y redes contribuyen a arquitecturas de computadoras distribuidas y cooperativas.
La figura 3 muestra el modelo de datos que presenta dos esquemas de información, lo que permite ampliar cada uno de ellos e incluso su reutilización en otros proyectos. Por un lado, se presenta el esquema correspondiente a Seguridad, para el tratamiento de perfiles, roles, usuarios, y administración de auditorías de información en general. Este esquema es el que permite poner disponible la capa de aplicación al usuario. Por otro lado se presenta el modelo propio del dominio del problema. Entre algunas de las virtudes que presenta se encuentran, la reutilización de arquitectura en distintos proyectos, múltiples matrices por proyectos y la posibilidad de contar con diversos planes de pruebas ejecutados sobre un mismo requerimiento.

Fig. 2: Diagrama de Casos de uso de la aplicación

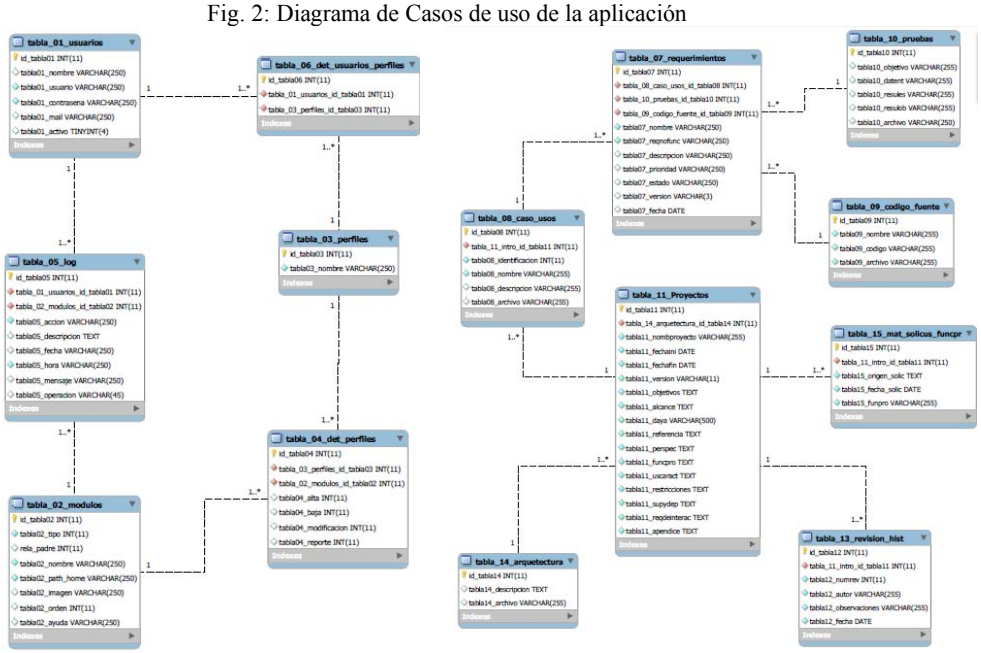


Fig. 3: Modelo de datos

2.1 Descripción de la herramienta

Pensada para gestionar trazabilidad, la herramienta permite incorporar información correspondiente a solicitudes de usuarios desde la primera etapa de un proyecto, como así también realizar la gestión de cambios y configuraciones de manera ordenada contribuyendo a la liberación de productos con un alto grado de control sobre las diferentes versiones de software, siguiendo las recomendación de buenas prácticas en Administración de Servicios de TI (ITIL Versión3).

Contempla todas las etapas del ciclo de desarrollo y contribuye a la comunicación entre el equipo de desarrollo y clientes, desde la posibilidad de contar con una visualización explícita de las solicitudes realizadas y posteriormente materializadas como funcionalidad en un producto.

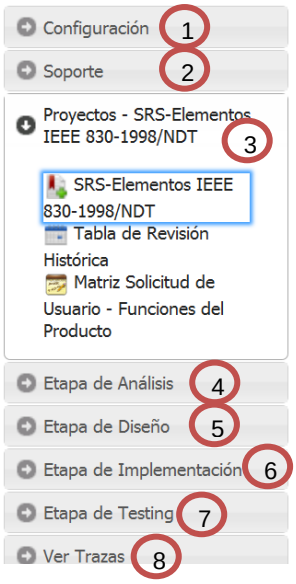
Permite establecer las relaciones entre los componentes del software, como por ejemplo: cuál es el origen de una especificación de análisis, diagramas construidos en cualquier herramienta CASE o similar, casos de pruebas, objetos de código fuente (compilado construidos, etc.), de manera que se pueda seguir la vida de un requerimiento en ambas direcciones, hacia delante y hacia atrás, es decir, desde su origen y especificación hasta su implementación.

La aplicación se desarrolló utilizando herramientas open source, tales como, JavaScript, jQuery,PHP,WAMP, CSS, Aptana Studio 3 (herramienta profesional para desarrollo de código abierto para la web), HTML 5, Canvas (nuevo componente de HTML 5, que permite dibujar por medio de un API), Firebug y MySQL Workbench.

La figura 4 muestra la interfaz inicial de la aplicación y las distintas opciones del menú principal.

Fig. 4: Interfaz inicial

El menú de la figura 5 muestra las distintas opciones que ofrece la aplicación.



1. **Configuración:** Permite configurar los módulos que conforman las distintas opciones del menú.
2. **Soporte**: Permite administrar usuarios, perfiles, carpetas de copias y recuperación de información.
3. **Proyectos - SRS-Elementos IEEE 830-1998/NDT**: Comprende las opciones que presentan diferentes aspectos de la información generada en el proyecto:
   3.1. SRS-Elementos IEEE 830-1998/NDT: Se completan los distintos ítems de la plantilla IEEE 830-1998/NDT.
   3.2. Tabla de revisión histórica: Permite obtener información de auditoría sobre las interacciones que han realizado los distintos usuarios.
   3.3. Matriz Solicitud de usuario – Funciones del Producto: Permite conocer el origen de las funcionalidades del proyecto.

Fig. 5: Opciones de la aplicación

Dentro de la opción 3, SRS-Elementos IEEE 830-1998/NDT, al incorporar un proyecto, se ingresan todos los ítems establecidos en la plantilla IEEE reformulada para su adecuación a desarrollo de aplicaciones web. La figura 6 muestra la interfaz prevista para esta opción.

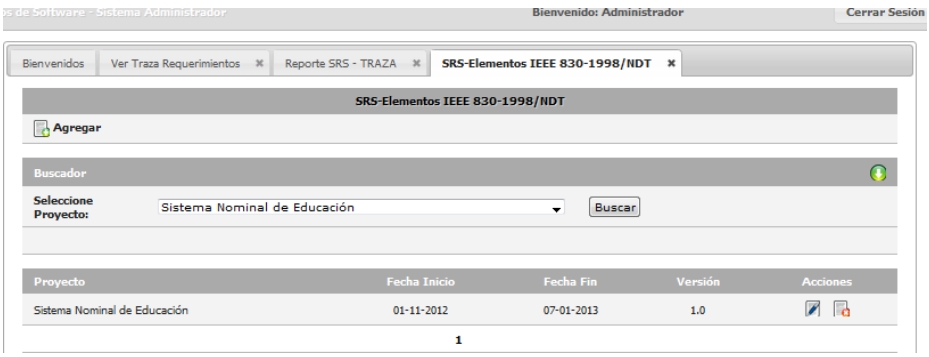


Fig. 6: SRS-Elementos IEEE 830-1998/NDT

Una vez seleccionado el proyecto, se completan los datos requeridos (ver figura 7).

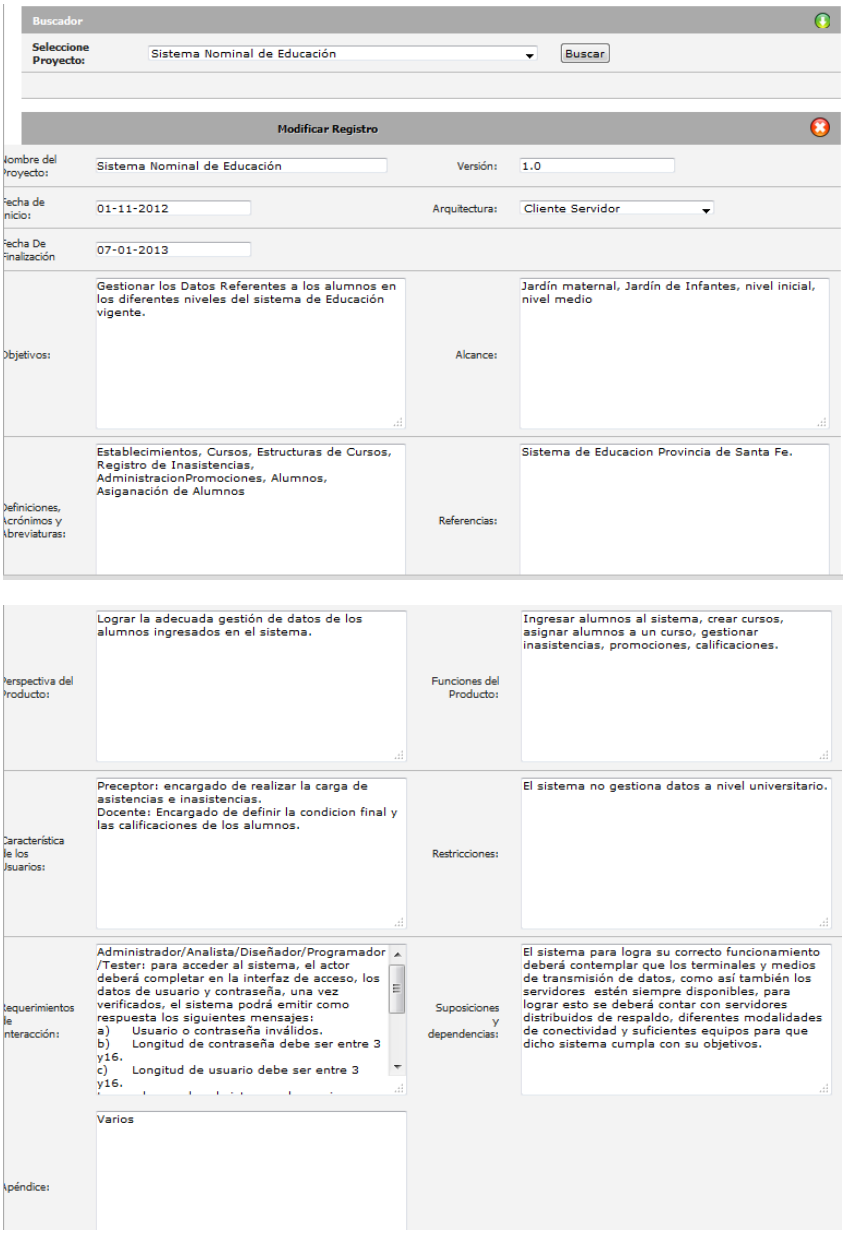
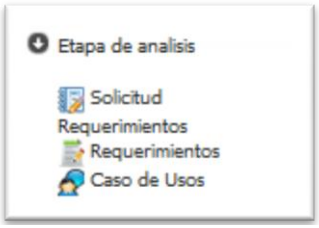


Fig. 7: SRS-Elementos IEEE 830-1998/NDT

A continuación, se describen los otros componentes del menú principal:



4. Etapa de análisis: permite la incorporación de diagramas y descripciones sobre documentación de análisis, provenientes de distintas fuentes y formatos. Las opciones ofrecidas se observan en la figura 8, cada una genera datos necesarios para la trazabilidad.

Fig. 8: Etapa de análisis

La figura 9 muestra las solicitudes iniciadas y la opción de agregar una solicitud.

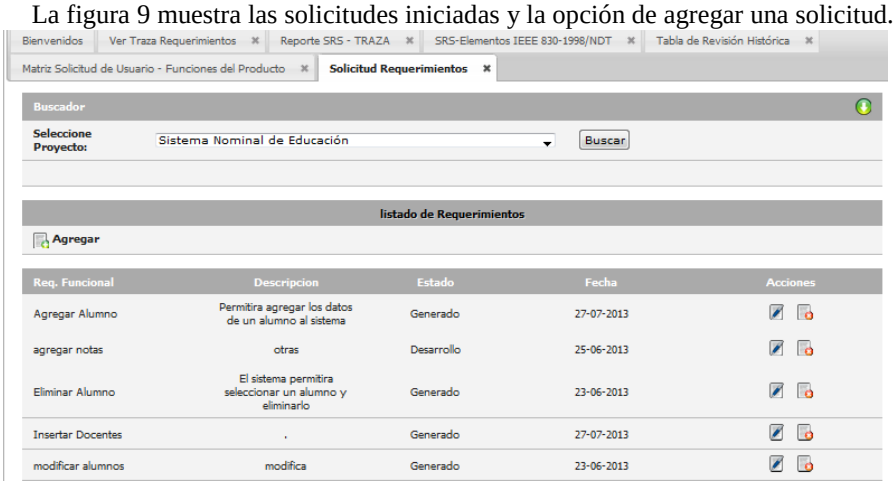


Fig. 9: Administración de requerimientos en la etapa de análisis

También es posible ver los distintos casos de uso, que incluyen una breve descripción, facilitando su comprensión (figura 10).

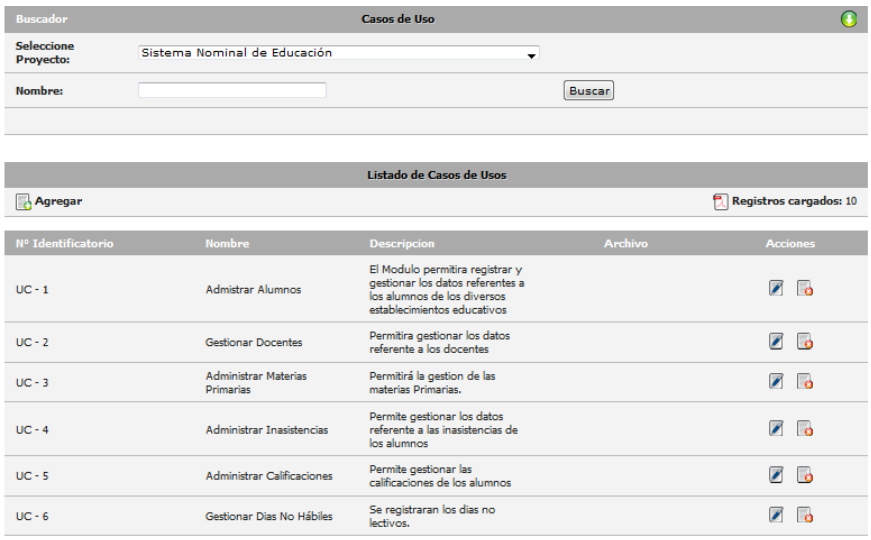


Fig. 10: Casos de uso en la etapa de análisis

5. **Etapa de diseño:** permite la incorporación de documentación relacionada a la arquitectura del proyecto.

6. **Etapa de implementación:** permite importar objetos construidos para soportar la funcionalidad, por ejemplo, asociar un componente compilado o incorporar el código.

7. **Etapa de testing:** corresponde a la vinculación de todo elemento construido para la verificación y/o validación de un componente de software, en cada una de sus etapas.

8. **Ver trazas:** presenta la información de las relaciones generadas en la vida de un requerimiento.

Por ejemplo, para el requerimiento Agregar Alumno, los mecanismos de trazabilidad previstos permiten visualizar toda la información relacionada (figura 11).

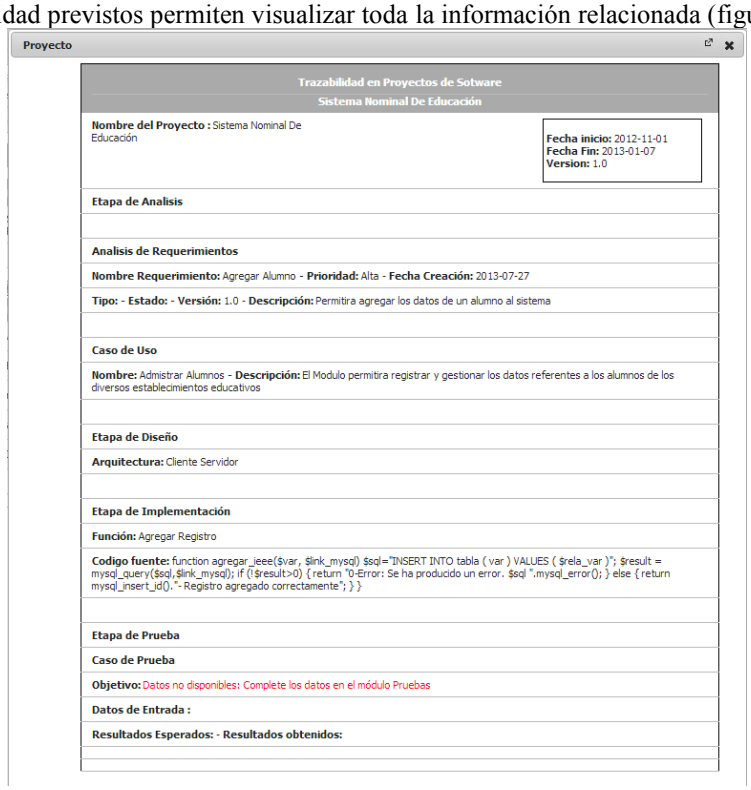


Fig. 11. Información relacionada al requerimiento Agregar Alumno

3 Conclusiones

Se construyó una aplicación web que permite administrar proyectos de software y realizar el seguimiento de los requerimientos en cada una de las fases del desarrollo, brindando información sobre el origen de cada cambio que ha sufrido una versión del software.

Como fortalezas de la herramienta se destacan la capacidad de seguir el ciclo de vida de un requerimiento, desde el origen de la especificación hasta la prueba del mismo, lo cual permite estimar el impacto de un cambio de requerimiento, la capacidad de descubrir dependencias y conflictos entre requerimientos, y la capacidad de mejorar la comprensión del sistema en su totalidad, características que contribuyen a facilitar el desarrollo y mantenimiento del mismo.

Como trabajo futuro se plantea validar la herramienta en ambientes de producción reales, tales como áreas de Sistemas de las organizaciones o empresas de software, para obtener una retroalimentación que permita mejorar la solución que se propone.

Referencias

1. Anaya R., Tabares M. S., Arango F.; "Una revisión de modelos y semánticas para la trazabilidad de requerimientos"; Revista EIA, ISSN 1794-1237 Nº 6, p. 33-42. 2006
2. Olsina, L. "Metodología cualitativa para la evaluación y comparación de la calidad de sitios web". Ph. Tesis.Facultad de Ciencias Exactas. Universidad de La Pampa. Argentina. 1999.
3. Escalona, M.J. "Metodología para el desarrollo de sistemas de información global: análisis comparativo y propuesta". Universidad de Sevilla. 2002.
4. Escalona, M.J., Mejías, M., Torres, J. "Methodologies to develop web information systems and comparative analysis". UPGRADE.TVol. III, No. 3, Junio 2002.
5. Escalona, M.J., Torres, J., Mejías, M. "Requirements capture workflow in Global Information Systems". Proceedings of OOIS.Springer-Verlag. Montpellier, France. 2002.
6. Escalona, M. and Koch, N., Requirements engineering for Web Applications: a comparative study. Journal of Web Engineering, 2004. 2: p. 193-212.
7. Escalona, M. J. and Koch, N. "Metamodeling the Requirements of Web Systems". In Proceedings of the Web Information Systems and Technologies (Setubal, Portugal, 2006).
8. Molina, F., Pardillo, J. and Toval, A., "Modelling web-based systems requirements using WRM". Web Information Systems Engineering–WISE 2008 Workshops, 2008. p. 122-131.
9. Aguilar, J. A., Garrigos, I., Mazon, J.-N. and Trujillo, J. "Web Engineering Approaches For Requirement Analysis - A Systematic Literature Review". 6th Web Information Systems and Technologies (WEBIST). 2010. Valencia, Spain.
10. Ramesh B. "Factors influencing requirements traceability practice". Communication of the ACM. Vol. 41, No. 12, pp. 37-44, December 1998.
11. Nicolás, J. and Toval, A., "On the generation of requirements specifications from software engineering models: A systematic literature review". Information and Software Technologies, 2009. 51(9): p. 1291-1307
12. INTECO, "Guía práctica de gestión de requisitos", [página web en línea]. Disponible en Internet http://www.inteco.es/file/NRDmviQoTbI_jZcyjTYRlw. Consulta: 19/07/2013.
13. Ferraro, M.; Medina, Y.; Dapozo, G.; Estayno, M. "Especificación y trazabilidad de requerimientos en el desarrollo de aplicaciones web". II Jornadas de Investigación en Ingeniería del NEA y Países Limítrofes. Facultad Regional Resistencia. Universidad Tecnológica Nacional. ISBN: 978–950–42–0142–7. Resistencia. Chaco. 2012.

IV WORKSHOP ASPECTOS TEÓRICOS DE CIENCIA DE LA COMPUTACIÓN

- WATCC -

CACIC2013

IV WORKSHOP ASPECTOS TEÓRICOS DE CIENCIA DE LA COMPUTACIÓN

- WATCC -

ID	Trabajo	Autores
5756	**A Complexity Lower Bound Based On Software Engineering Concepts**	Andrés Rojas Paredes (UBA)
5822	**On Aggregation Process in Linguistic Decision Making Framework**	M. Giménez, S. Gramajo (UTN)

A Complexity Lower Bound Based On Software Engineering Concepts

Andrés Rojas Paredes

Universidad de Buenos Aires,
Facultad de Ciencias Exactas y Naturales, Departamento de Computación.
Pabellón 1, Ciudad Universitaria, Buenos Aires, Argentina.
arojas@dc.uba.ar

Abstract. We consider the problem of polynomial equation solving also known as quantifier elimination in Effective Algebraic Geometry. The complexity of the first elimination algorithms were double exponential, but a considerable progress was carried out when the polynomials were represented by arithmetic circuits evaluating them. This representation improves the complexity to pseudo–polynomial time.

The question is whether the actual asymptotic complexity of circuit–based elimination algorithms may be improved. The answer is no when elimination algorithms are constructed according to well known software engineering rules, namely applying information hiding and taking into account non–functional requirements. These assumptions allows to prove a complexity lower bound which constitutes a mathematically certified non–functional requirement trade–off and a surprising connection between Software Engineering and the theoretical fields of Algebraic Geometry and Computational Complexity Theory.

Keywords: Non-functional requirement trade–off, information hiding, arithmetic circuit, complexity lower bound, polynomial equation solving, quantifier elimination in algebraic geometry

1 Introduction

The main issue of this paper is to describe the Software Engineering aspects of the mathematical computation model introduced in [9]. This model captures the notion of a circuit–based elimination algorithm in order to solve a thirty years old problem in algebraic complexity theory (see e.g. [8], [10]): in *arithmetic circuit–based* effective elimination theory the elimination of a single existential quantifier block in the first order theory of algebraically closed fields of characteristic zero is *intrinsically hard* (i.e. it has an *exponential* complexity lower bound). This conclusion may also be expressed in terms of a trade–off between two non–functional requirements: on one hand we have a complexity requirement and on the other a property of mathematical functions called *geometrical robustness*. This complexity lower bound in terms of software engineering concepts appears

for first time in [5] in the context of polynomial interpolation. In this work we study a more general case in the context of quantifier elimination.

Complexity lower bounds are undoubtedly theoretical research. But there is also a practical aim behind that. Consider the process in software design where a software architecture is developed in order to solve a certain computational problem. Assume also that one of the non–functional requirements of the software design project consists of a restriction on the run time computational complexity of the program which is going to be developed (this was the case during the implementation of the polynomial equation solver Kronecker by G. Lecerf, see [11]). Our practical aim is to provide the software engineer with an efficient tool which allows him to answer the question whether his software design process is entering at some moment in conflict with the given complexity requirement. If this is the case, the software engineer will be able to change at this early stage his design and may look for an alternative software architecture. The following example illustrates this description.

Example 1 (Finite Set). Suppose that our task is to implement a finite set S of cardinality n, e.g. a subset of the natural numbers $\mathbb{N}$, and that we have to satisfy the requirement that membership to the finite set S is decided using only $O(\log n)$ comparisons. If the set S is implemented by an unordered array, we will be unable to satisfy our complexity requirement. So we are forced to think in alternative implementations of the abstract concept of a finite set, e.g. by ordered arrays, special trees or any other data type which is well suited for our task.

Example 1 represents a case where it may be impossible to satisfy a given complexity requirement by means of a previously fixed software architecture. Our aim is to formalise such impossibility by means of a complexity lower bound which is usually difficult to infer when the number of components of the system under consideration is large or when the predicate to decide or the function to compute becomes more sophisticated like in polynomial equation solving. This leads to the idea to fix in advance only a small selection of architectural features, e.g. the abstraction levels or part of the language of our system (not the algorithms themselves). The computation model we are going to explain in following sections takes into account these considerations.

This work is organised as follows: in Section 2 we introduce quantifier elimination as the subject of our complexity studies and the algorithmic approach which is based on the transformation of arithmetic circuits. In Section 3 we describe the tool used to obtain the announced complexity lower bound. Our tool is a computation model which captures the notion of non–functional requirement in circuit–based elimination algorithms. Finally we present the new result in this work: we make the following question: What does it happen if our algorithms are not circuit–based and we found a representation which is more efficient than circuits? The answer is that our complexity results are valid for arbitrary continuous representations if the algorithms follow the principle of *information hiding*. We illustrate this conclusion with a relevant example from the theory of Abstract Data Types (see, e.g. [13] and [12]).

In the rest of the paper we shall use notions and notations from algebraic geometry and algebraic complexity theory which are all standard (see for example [14] and [3]).

2 Quantifier elimination and its implementation

2.1 Quantifier Elimination

We start with the subject of our complexity studies. The subject is *quantifier elimination in the particular case of elementary algebraic geometry over* $\mathbb{C}$. Let Φ be an existentially quantified formula. In general terms, the quantifier elimination problem consists in obtaining a quantifier free formula Ψ which is logically equivalent to Φ (this means that Ψ and Φ define the same set). In the particular case of elementary algebraic geometry over $\mathbb{C}$, the formulas Φ and Ψ are composed by polynomial equations. In this context we are going to consider exclusively the polynomials of these equations.

Let n and r be natural numbers. Let T, $U := (U_1, \ldots, U_r)$ be *parameters* and $X := (X_1, \ldots, X_n)$ be *variables* subject to quantification. We focus our attention to polynomials $G_1(X), \ldots, G_n(X)$ and $H(T, U, X)$ which belong to $\mathbb{C}[X]$ and to $\mathbb{C}[T, U, X]$ respectively. These polynomials constitute a so called *Flat Family of Elimination Problems* given by the polynomial equation system $G_1 = 0, \ldots, G_n = 0$ and the polynomial H (see, e.g. [4] and [9] for details). In general terms this system represents the quantified formula $\Phi : (\exists X_1)(\exists X_n)(G_1 = 0 \wedge \ldots \wedge G_r = 0 \wedge Y - H = 0)$.

On the other hand, there exists a polynomial $F \in \mathbb{C}[T, U, Y]$ of minimal degree, called the associated *Elimination Polynomial*, such that the equation $F = 0$ represents a quantifier–free formula Ψ which is equivalent to Φ.

Thus, we arrive to a functional requirement where the flat family of elimination problems given by $G_1 = 0, \ldots, G_n = 0$ and H becomes transformed into the elimination polynomial F. This transformation is carried out by a mathematical function f as Fig. 1 illustrates.

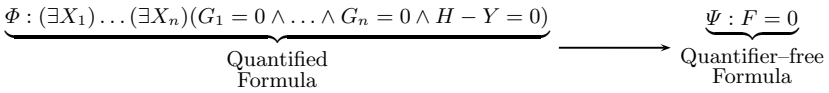


Fig. 1: Quantifier elimination problem.

At this abstract level we do not know, for example, how the polynomials are implemented in the computer. We define now these implementation details.

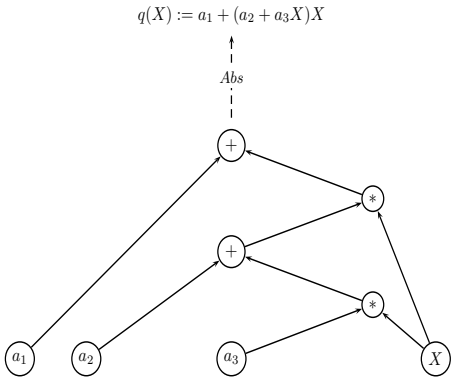
An implementation option is to represent polynomials by their coefficients. Unfortunately the coefficient representation in some elimination polynomials

may conduce to complexity blow ups, e.g. the Pochhammer polynomial

$$\prod_{0 \leq j < 2^n} (Y - j)$$

which has 2^n terms (see [7] for open questions in complexity theory related to this polynomial). This circumstance suggests to represent in elimination algorithms polynomials not by their coefficients but by arithmetic circuits. This idea became fully realised by the "Kronecker" algorithm for the resolution of polynomial equation systems over algebraically closed fields. The algorithm was anticipated in [6] and implemented in a software package of identical name (see [11]). The following example illustrates the notion of arithmetic circuit.

Fig. 2: Arithmetic circuit and Horner scheme.



Example 2 (Horner scheme). Let a_1, a_2, a_3 be constants and X be a variable. Consider the polynomial $p(X) = a_1 + a_2X + a_3X^2$ and the Horner scheme of this polynomial which is $q(X) = a_1 + (a_2 + a_3X)X$. From this scheme we have a directed acyclic graph where each node is an arithmetic operation $+, *$, a constant a_1, a_2, a_3 or a variable X. This arithmetic circuit is a concrete object implementing the abstract object $q(X)$. Fig. 2 illustrates the relation between $q(X)$ and its implementing circuit by means of an abstraction function *Abs*.

2.2 Implementation of quantifier elimination

To understand the role of arithmetic circuits in elimination algorithms we fix the notion of polynomials in terms of abstract data types and classes implementing them. Here we follow the terminology in [13].

Suppose that we have an abstract data type specification for polynomials in terms of query and creator functions (observers and constructors in the terminology of [12]). Thus the elimination problem of Fig. 1 may be expressed as a specification in terms of abstract data types.

Consider now the classes implementing the abstract data type of polynomials. We have a class for polynomials and a class for circuits. The connection between these two classes is that the class of circuits is a private part of the class of polynomials. This private part is used to implement the interface of the class of polynomials in terms circuits. In this context polynomials are encapsulated circuits which are mapped into instances of the abstract data type of polynomials by an abstraction function *Abs*.

Now recall our functional requirement: transform en elimination problem given by polynomials $G_1, \ldots, G_n$ and H into an elimination polynomial F. Since

polynomials become implemented by circuits, an elimination algorithm works directly with circuits taking care of satisfy class invariants and the abstraction function *Abs*. In this sense, an elimination algorithm $\mathcal{A}$ transforms an input circuit β representing $G_1, \ldots, G_n, H$ into an output circuit γ representing the elimination polynomial F as Fig. 3 illustrates.

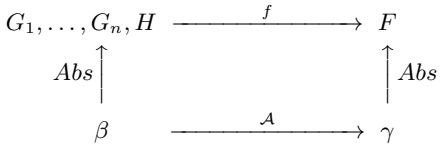


Fig. 3: Elimination problem and its implementation.

The transformation of β into γ is carried out by means of circuit operations, e.g. join of circuits which mimics the composition of functions (see also union of circuits and recursive routine in [9]). If we require the algorithm $\mathcal{A}$ to be *branching parsimonious* (see Section 3.1 below), then $\mathcal{A}$ captures all known circuit–based elimination algorithms including the polynomial equation solver Kronecker.

At this point the question is how we measure the complexity of algorithm $\mathcal{A}$. We shall mainly be concerned with the *size* of the output circuit γ. Here we refer with "size" to the number of internal nodes which count for the given complexity measure. Our basic complexity measure is the *non–scalar* one (also called *Ostrowski measure*) over the ground field $\mathbb{C}$. This means that we count, at unit costs, only essential multiplications and divisions (see [3] for details).

3 Software engineering–based approaches to complexity lower bounds

3.1 A circuit–based computation model

The polynomials $G_1, \ldots, G_n, H$ and F described before belong to mathematical structures $\mathbb{C}[X]$, $\mathbb{C}[T, U, X]$ and $\mathbb{C}[T, U, Y]$ respectively. In these mathematical structures polynomials have a natural property called *geometrical robustness* which interpreted as a non–functional requirement constitutes a key ingredient in our complexity result (see Theorem 1 below). This property is invisible if we only consider abstract data type specifications in the sense of [13]. Thus, in order to include geometrical robustness in the specification of elimination problems we model the notion of abstract data type of polynomials with the corresponding mathematical structure and we call this structure an abstract data type. For example, the polynomials $G_1, \ldots, G_n, H$ will be instances of the abstract data type $\mathcal{O} \subset \mathbb{C}[T, U, X]$ and F will be an instance of the abstract

data type $\mathcal{O}^* \subset \mathbb{C}[T, U, Y]$ and the elimination problem will be specified by a geometrically robust map $f : \mathcal{O} \to \mathcal{O}^*$.

Geometrical robustness The map f is a function (mathematical application) which we require to be constructible, i.e. definable by a boolean combination of polynomial equations. The mapping is called geometrically robust if it is continuous (see [9] and [5] for an algebraic characterisation of robustness). Since geometrical robustness is a property belonging to the specification level of our elimination task, we have to describe how this non–functional property is realised by the circuit–based algorithms implementing the elimination.

Branching parsimoniousness The intuitive meaning of geometrical robustness is reflected by the algorithmic notion of *branching parsimoniousness*. We call an algorithm branching parsimonious if it avoids unnecessary branchings. We may restrict branchings by means of only considering division–free circuits, or circuits where divisions by zero were replaced by suitable limits and divisions may only involve parameter nodes (nodes without variables). In this sense our circuits are *essentially division–free* and will be called *robust* if all intermediate results (functions represented by each node) are geometrically robust.

The notion of branching parsimoniousness as a tactic In the context of software architecture, the satisfaction of quality attributes requires techniques which are called *tactics*. For example, a system is easily modified when it is structured, modularised and well documented. A tactic is, according to [2], a design decision that influences the control of a quality attribute response. Considering this definition we may describe branching parsimoniousness as a tactic for elimination algorithms. We require an algorithm to be branching parsimonious in order to achieve the non–functional requirement of geometrical robustness. In this sense we say that branching parsimoniousness is a tactic to achieve geometrical robustness. For example, the reader may identify branching parsimoniousness with *modularity* which is a tactic to achieve the modifiability quality attribute.

Now recall our elimination algorithm $\mathcal{A}$ in Fig. 3 which transform the circuit β (representing $G_1, \ldots, G_n, H$) into circuit γ (representing F). The elimination algorithm $\mathcal{A}$ implements the additional property of geometrical robustness if we require $\mathcal{A}$ to be banching parsimonious.

Thus in the input we have an essentially division–free, robust parameterized arithmetic circuit β of size $O(n)$ with basic parameters T, $U := U_1, \ldots, U_n$ and input $X := X_1, \ldots, X_n$ which computes polynomials $G_1, \ldots, G_n \in \mathbb{C}[X]$ and $H \in \mathbb{C}[T, U, X]$ constituting a flat family of zero–dimensional elimination problems with associated elimination polynomial $F \in \mathbb{C}[T, U, Y]$.

The branching parsimoniousness allows to affirm that each circuit operation gives as result a robust circuit. Thus we conclude that the property of geometrical robustness is transmitted from the input β to the output γ. Then $\gamma := \mathcal{A}(\beta)$ is an essentially division–free, robust parameterized arithmetic circuit with basic parameters $T, U_1, \ldots, U_n$ and input Y representing the elimination polynomial F.

These notations and assumptions, in particular the property of robustness in the output γ, allows to conclude the following theorem.

Theorem 1 ([9], Theorem 10). *The circuit γ has, as ordinary arithmetic circuit over $\mathbb{C}$, non–scalar size at least $\Omega(2^n)$.*

Theorem 1 corresponds to circuit–based algorithms, now we ask what does it happen if we found a representation which is more efficient than arithmetic circuits? We argue that Information Hiding–based algorithms have the same complexity status. This implies that our complexity results are valid for arbitrary continuous representations. This is part of future work but we give preliminary results in the following section.

3.2 Towards an Information Hiding–based computation model

Since polynomials $G_1, \ldots, G_n, H$ and F are objects belonging to suitable abstract data types, we may define the function f of Fig. 3 in terms of query and creator functions (observers and constructors) of the given abstract data type specification obtaining a transformation which does not involve circuits directly because they become encapsulated. To illustrate this kind of transformation consider the following example.

Example 3. Suppose a case where f is the identity function of binary trees. In this context let us consider the following abstract functions of the corresponding abstract data type specification: $root()$, $left()$, $right()$ and $isNil?()$ as query functions (observers) and $bin()$ and $nil()$ as creator functions (constructors). Then, we propose the following definition for f:

$$f(X) = \begin{cases} nil() & \text{if } isNil?(X) \\ bin(root(X), id(left(X)), id(right(X))) & \text{otherwise} \end{cases} \quad (1)$$

This specification of function f may be implemented in such a way that, at an abstract level the implementation is the identity function of binary trees, whereas at a concrete level the implementation is a transformation of the representation of binary trees (compare this with the transformation of circuits in elimination algorithm $\mathcal{A}$). This hidden transformation is carried out by the classes implementing the abstract data type of binary trees where the implementation of f may be called `f`. Notice that we write the implementation in `verbatim` font in order to distinguish the difference with abstract data type expressions which we write in *cursive* font.

Let `Tree<E>` be a class implementing the abstract data type of binary trees. Let `Tree1<E>` and `Tree2<E>` be subclasses of class `Tree<E>` with the following property: `Tree1<E>` implements trees as arrays (the internal representation of trees is given by arrays) and `Tree2<E>` implements trees as nested nodes. Let `root`, `left`, `right` and `isNil` be routines in the class `Tree<E>` implementing the corresponding query functions (observers) in the abstract data type specification.

Let `p1` be a variable of type `E` and `p2` y `p3` be variables of type `Tree2<E>`, then `Tree2<E>()` and `Tree2<E>(p1,p2,p3)` are constructors of the class `Tree2<E>` implementing the creator functions $nil()$ and $bin()$ respectively. Then the implementation in java code is as follows:

```
Tree<E> f(Tree<E> t){
        if(t.isNil()) return new Tree2<E>();
        else return new Tree2<E>(t.root(),
                (Tree2<E>) f(t.left()),
                (Tree2<E>) f(t.right()) );
}
```
(2)

Notice that the effective transformation of the representation is carried out when an instance of `Tree1<E>` is passed as parameter and the constructor of the other class is applied, say the constructor of `Tree2<E>`.

Equation 2 illustrates the definition of an algorithm in terms of observers and constructors. In the case of elimination problems such an algorithm has a similar structure but we do not exhibit an example here. This is left for a future work (see [1]) where the notion of information hiding is modelled in full detail. Such a model allows to conclude the following:

- if the complexity measure is given by the number of parameters instead of the size of circuits, we obtain an exponential complexity lower bound for this quantity which implies the result in Theorem 1,
- this allows to conclude that elimination algorithms programmed with information hiding, i.e. hiding the circuits or any other representation of polynomials, have the same complexity status.

Final comments The circuit–based computation model described in Section 3.1 corresponds to the tool for the software engineer we described at the introduction. Of course this model cannot be applied to any software project since it is restricted to the particular case of elimination. However, it gives the key ingredients for the definition of a computation model suitable for complexity questions where another non–functional requirement must be considered.

On the other hand, our description of an Information Hiding–based computation model in Section 3.2 constitutes an stronger result which together with Theorem 1, allows to conclude that the Kronecker algorithm is asymptotically optimal. This suggest that the Kronecker is a good option to use in applications of scientific computing where polynomial equation solving is needed.

Finally, a computation model which captures algorithms constructed in a professional way, namely applying software engineering concepts, in combination with the complexity lower bound obtained in Section 3.1 allows to conclude the following idea which we repeat from [9]: neither mathematicians nor software engineers, nor a combination of them will ever produce a practically satisfactory, *generalistic* software for elimination tasks in Algebraic Geometry. This is a job for *hackers* which may find for *particular* elimination problems *specific* efficient solutions.

Acknowledgements *The author thanks Joos Heintz for his insistent encouragement to finish this work and Pablo Barenbaum, Gastón Bengolea Monzón, Mariano Cerrutti, Carlos Lopez Pombo, Hvara Ocar and Alejandro Scherz, Universidad de Buenos Aires, for discussions about the topic of this paper and/or comments and ideas on earlier drafts.*

References

1. Bank, B., Heintz, J., Pardo, L.M., Rojas Paredes, A.: Quiz games: A new approach to information hiding based algorithms in scientific computing, manuscript Universidad de Buenos Aires (2013)
2. Bass, L., Clements, P., Kazman, R.: Software Architecture in Practice. Addison–Wesley, Boston, MA, 2. edn. (2003)
3. Bürgisser, P., Clausen, M., Shokrollahi, M.A.: Algebraic Complexity Theory. Grundlehren der mathematischen Wissenschaften, vol. 315. Springer Verlag (1997)
4. Castro, D., Giusti, M., Heintz, J., Matera, G., Pardo, L.M.: The hardness of polynomial equation solving. Foundations of Computational Mathematics 3(4), 347–420 (2003)
5. Giménez, N., Heintz, J., Matera, G., Solernó, P.: Lower complexity bounds for interpolation algorithms. Journal of Complexity 27, 151–187 (2011)
6. Giusti, M., Heintz, J., Morais, J., Morgenstern, J., Pardo, L.: Straight-line programs in geometric elimination theory. Journal of Pure and Applied Algebra 124, 101–146 (1998)
7. Heintz, J., Morgenstern, J.: On the intrinsic complexity of elimination theory. Journal of Complexity 9, 471–498 (1993)
8. Heintz, J., Sieveking, M.: Absolute primality of polynomials is decidable in random polynomial time in the number of variables. Automata, languages and programming (Akko, 1981). Lecture Notes in Computer Science 115, 16–28 (1981)
9. Heintz, J., Kuijpers, B., Rojas Paredes, A.: Software engineering and complexity in effective algebraic geometry. Journal of Complexity (2012)
10. Kaltofen, E.: Greatest common divisors of polynomials given by straight–line programs. J. Assoc. Comput. Mach. 35(1), 231–264 (1988)
11. Lecerf, G.: Kronecker: a Magma package for polynomial system solving. Web page. `http://lecerf.perso.math.cnrs.fr/software/kronecker/index.html`
12. Liskov, B., Guttag, J.: Program development in Java: Specification, and Object–Oriented Design. Addison-Wesley, 3. edn. (2001)
13. Meyer, B.: Object-Oriented Software Construction. Prentice-Hall, 2. edn. (2000)
14. Shafarevich, I.R.: Basic algebraic geometry: Varieties in projective space. Springer, Berlin Heidelberg, New York (1994)

On Aggregation Process in Linguistic Decision Making Framework

M. Gimenez, S. Gramajo

Artificial Intelligence Research Group. National Technological University,
French 414, Resistencia, 3500, Argentina
manuelego1@gmail.com, sergio@frre.utn.edu.ar

Abstract. When solving a problem, human beings must face situations in which they should choose among different alternatives by means of reasoning and mental processes. Many of these decision problems are under uncertain environments including vague, imprecise and subjective information that is usually modeled by fuzzy linguistic approach. This approach uses linguistic information or natural language words and its relation to mental reasoning processes of the experts when expressing their assessments. In a decision process multiple criteria can be evaluated which involving multiple experts with different degrees of knowledge. Such process can be modeled by using Multi-granular Linguistic Information (MGLI) and Computing with Words (CW) processes to solve the related decision problems. Once decision makers (experts) provided their opinions, it is necessary to combine all these opinions to obtain a single overall result that can be interpreted. An aggregation operator allows accomplishing this objective calculating a global value in different ways. In this paper we study the use of aggregation operators in multi-criteria decision-making processes comparing them and obtaining conclusions about their use in our framework. Furthermore, we propose a new aggregation operator taking into account the criteria importance to evaluate the alternatives, and then an illustrative example shows its outcomes.

Keywords: Multi-granular Linguistic Information, Computing with Words, Aggregation operator, Decision Making.

1 Introduction

The decision making is a day-to-day activity for human beings. The multiple facets of real world decision problems are well addressed by Multi-Criteria Decision Making (MCDM) [1]. The crucial point of interest within the MCDM is the analysis and the modeling of the multiple decision makers' preferences giving rise to Multi-Expert Decision Making (MEDM) [2].

In many situations, context involves vague and probably incomplete information. In these cases, information cannot be assessed precisely in a quantitative form; experts may feel more comfortable employing other approaches. To overcome this problem, information is normally modeled by using a fuzzy linguistic approach [3][4][5]

allowing the experts to express their opinions with words rather than numbers (e.g. when evaluating the comfort or design of a car, terms like good, medium, bad can be used). Therefore, the fuzzy linguistic approach is a technique that represents qualitative information as linguistic values by means of linguistic variables [3], that is, variables whose values are no numbers but words or sentences in a natural language. Each linguistic value is characterized by its syntax (label) and semantic (meaning). The label is a word or a sentence belonging to a linguistic term set and the meaning is a fuzzy subset in a universe of discourse. The concept of linguistic variables provides an estimated measure since words are less precise than numbers. This is more effective because the experts may feel more comfortable using words they really know and understand in accordance with the context of use of these words. Also, when offering different expression domains or different linguistic term sets (multi-granular information) to the experts, this solution would be suitable to adjust the degree of experience of each one [6][7].

An important aspect of the MCDM is the aggregation process. In order to obtain a unique final result, the assessments of each expert involved must be taken in account. An aggregation operator allows accomplishing this objective calculating a global value. The aggregation is the operation that transforms a set of elements, such as individual opinions on a set of alternatives, into a single element that is representative of the whole. Different ways of carrying out the combination of preferences have led many authors to study and design different aggregation operators. Depending on the problem different types of aggregation operators can be used.

In this paper, we focus on the aggregation process when dealing with complex decisions under uncertainty using decision analysis process. We will study the results of applying different aggregation operators on the same decision problem in order to obtain relevant conclusions about their use in complex decision systems.

This paper is organized as follows. Section 2 reviews basic concepts about linguistic background that will be used to model uncertain information and multi-granular information in our framework. Section 3 presents the phases in order to analyze decisions, with special emphasis on aggregation process. Then, section 4 proposes an example of use on investment decisions in a company. Finally, section 5 shows some conclusions.

2 Preliminaries

Normally the decision analysis depends highly on subjective, vague and ill-structured information must have a model to manage this kind of information. Therefore, we consider the use of the fuzzy linguistic approach [3] to model and manage the inherent uncertainty in this kind of problems and the 2-tuple linguistic model to represent linguistic information [8]. Additionally, it is useful to manage multiple linguistic scales (multi-granular information) giving more flexibility to the different experts involved in the problem and, to manage this, we use Extended Linguistic Hierarchies (ELH) method. For this reason, in this section we review in short the concepts and

methods such as the fuzzy 2-tuple linguistic model, ELH and its computational method (aggregation process).

2.1 The 2-tuples linguistic model

When using linguistic information to solve a problem it is necessary the use of computing with words CW. The main limitation with this approach is the "loss of information" suffered in the most used computational techniques that implies the lack of precision in the final results. These computational models are: The semantic model [9] and the symbolic model [10]. In these two models an approximation process must be developed to express the result in the initial expression domain, here is when the information gets lost.

The 2-tuples linguistic model [11] is a representation model that overcomes the loss of information. It represents the linguistic information with a pair of values, that we call 2-tuple, composed by a linguistic term and a number.

Definition 1. The Symbolic Translation of a linguistic term $s_i \in S = \{s_0, ..., s_g\}$ is a numerical value assessed in $[-0.5, 0.5)$ that supports the "difference of information" between an amount of information $\beta \in [0, g]$ and the closest value in $\{0, ..., g\}$ that indicates the index of the closest linguistic term in $S(s_i)$, being $[0, g]$ the interval of granularity of S.

From this concept a new linguistic representation model was developed, which represents the linguistic information by means of a linguistic 2-tuple. It consists of a pair of values namely, $(s_i, \alpha) \in \overline{S} \equiv S \times [-0.5, 0.5)$, being $s_i \in S$ a linguistic term and $\alpha \in [-0.5, 0.5)$ a numerical value representing the symbolic translation. This representation model defined a set of transformation functions between numeric values and linguistic 2-tuples to facilitate linguistic computational processes.

Definition 2. Let $S = \{s_0, ..., s_g\}$ be a linguistic terms set and $\beta \in [0, g]$ a value supporting the result of a symbolic aggregation operation. The 2-tuple set associated with S is defined as $\overline{S} = S \times [-0.5, 0.5)$. A 2-tuple that expresses the equivalent information to β is then obtained as follow:

$$\Delta : [0, g] \to \overline{S}$$

$$\Delta(\beta) = (s_i, \alpha), with \begin{cases} s_i, & i = round(\beta) \\ \alpha = \beta - i, & \alpha \in [-0.5, 0, 5) \end{cases} \tag{1}$$

being round $(\cdot)$ the usual round operation, i the index of the closest label, s_i, to "β", and "α"the value of the symbolic translation.

It is noteworthy to point out that Δ is a one to one mapping and $\Delta^{-1} : \overline{S} \to [0, g]$ is defined by $\Delta^{-1}(s_i, \alpha) = i + \alpha$. In this way the 2-tuple of $\overline{S}$ is identified by a numerical value in the interval $[0, g]$.

Remark 1. The transformation of a linguistic term into a linguistic 2-tuples consists of adding value 0 as symbolic translation: $s_i \in S \Rightarrow (s_i, 0) \in \overline{S}$. On other hand, $\Delta(i) = (s_i, 0)$ and $\Delta^{-1}(s_i, 0) = i, \forall i \in \{0, 1, .., g\}$.

If $\beta = 3.25$ is the value representing the result of a symbolic aggregation operation on the set of labels,
$S = \{s_0 = Nothing, s_1 = VeryLow, s_2 = Low, s_3 = Mediums, s_4 = High, s_5 = VeryHigh, s_6 = Perfect\}$
, then the 2-tuple that expresses the equivalent information to β is $(medium, .25)$.

This model has a linguistic computational technique based on the functions Δ and Δ^{-1}, for a further detailed see Ref. [12].

2.2 Extended Linguistic Hierarchies:

A flexible expression domain with several linguistic scales is necessary to express the assessments for experts according to their degree of knowledge about the problem. Different approaches dealing with multi-granular linguistic information have been proposed. In this paper shall use the ELH [13] approach to model and manage multi-granular linguistic information because of its features of flexibility and accuracy in the processes of computing with words (CW) in multi-granular linguistic contexts. An ELH is a set of levels, where each level represents a linguistic term set with different granularity from the remaining levels of the ELH. Each level belongs to an ELH is denoted as $l(t, n(t))$ being t a number that indicates the level of the ELH and $n(t)$ the granularity of the terms set of the level t . To build an ELH have been proposed a set of extended hierarchical rules:

Rule 1: A finite set of levels, $l(t, n(t))$ with $t = 1, .., m$, that defines the multi-granular linguistic context required by experts to express their assessments are included.

Rule 2: to obtain an ELH a new level, $l(t^*, n(t^*))$ with $t^* = m + 1$, should be added. This new level must have the following granularity:

$$n(t^*) = (L.C.M.(n(1) - 1, ..., n(m) - 1)) + 1 \tag{2}$$

being L.C.M. the Least Common Multiple.

ELH building process then consists of two processes: i) It adds m linguistic scales used by the experts to express their information. And ii) then it adds the term set $l(t^*, n(t^*))$, with $t = m + 1$, according to Eq. (2). Therefore, the ELH is the union of all levels required by the experts plus the new level $l(t^*, n(t^*))$.

$$ELH = \bigcup_{t=1}^{t=m+1} (l(t, n(t)))$$

The use of multi-granular linguistic information makes the processes of CW more complex. ELH computational model needs to make a three-step process.

1. Unification phase. The multi-granular linguistic information is conducted into only one linguistic term set, that in ELH is always $S^{n(t^*)}$, by means of a transformation function $TF_b^a(\cdot)$:

Definition 3. Let $S^{n(a)} = \{s_0^{n(a)}, ..., s_{n(a)-1}^{n(a)}\}$ and $S^{n(b)} = \{s_0^{n(b)}, ..., s_{n(b)-1}^{n(b)}\}$ be two linguistic term sets, with $a \neq b$. The linguistic transformation function is defined as:

$$TF_b^a : \bar{S}^{n(a)} \rightarrow \bar{S}^{n(b)}$$

$$TF_b^a(s_j^{n(a)}, \alpha_j^{n(a)}) = \Delta_S\left(\frac{\Delta^{-1}(s_j^{n(a)}, \alpha^{n(a)}) \cdot (n(b)-1)}{n(a)-1}\right) = (s_k^{n(b)}, \alpha_k^{n(b)}) \quad (2)$$

2. Computational process. Once the information is expressed in only one expression domain $S^{n(t^*)}$, the computations are carried out by using the linguistic 2-tuple model.

3. Expressing results. In this step the results might be transformed into any level, t , of ELH in a precise way by using Eq. (3) to improve the understanding of the results if necessary.

Remark 2. In the processes of CW with information assessed in an ELH, the linguistic transformation function, TF_b^a , performed in the unification phase, a , might be any level in the set $\{t = 1, .., m\}$ and the computational processes are carried out in the level b that it is always the level t^* (See Eq. (3)).

It was proved in [13] that the transformation functions between linguistic terms in different levels of the Extended Linguistic Hierarchy are carried out without loss information.

2.3 Aggregation process:

Aggregation operators allow accomplishing a global value from a set of values in order to obtain a unique final value. Here we have analyzed four kinds of aggregation operators, Geometric Mean Aggregation Operator (GMAO), Arithmetic Mean Aggregation Operator (AMAO) Weighted Aggregation Operator (WAO). WAO is based on the weight of the experts (WAO) or criteria (WAOC).

Definition 6. GMAO. Let $((l_1, \alpha_1), .., (l_m, \alpha_m)) \in \bar{S}^m$ be a 2-tuples linguistic vector, geometric mean operator is defined as follows: $G : \bar{S}^m \rightarrow \bar{S}$

$$G : [((l_1, \alpha_1), .., (l_m, \alpha_m))] = \left[\prod_{i=1}^{m} \Delta^{-1}[(l_i, \alpha_i)]\right]^{\frac{1}{m}} = \left[\prod_{i=1}^{m} \beta_i\right]^{\frac{1}{m}} \quad (3)$$

Definition 5. AMAO: Let $((l_1, \alpha_1), .., (l_n, \alpha_n)) \in \bar{S}^n$ be a 2-tuples linguistic vector, arithmetic mean operator is defined as follows: $\overline{G} : \bar{S}^n \rightarrow \bar{S}$

$$\overline{G}[(l_1, \alpha_1), .., (l_n, \alpha_n)] = \Delta(\sum_{j=1}^{n} \frac{1}{n} \Delta^{-1}(r_j, \alpha_j)) = \Delta(\frac{1}{n} \sum_{j=1}^{n} \beta_j) \quad (4)$$

A rational assumption about the resolution process could be associating more importance to the experts who have more "knowledge" or "experience". These values can be interpreted as *importance degree*, *competence*, *knowledge* or *ability* of the experts. In addition some experts could have some difficulties in giving all their assessments due to lack of knowledge about part of the problem. Besides the use of different scales, the expert should be carried out in different way with weighted aggregation operator.

Definition 6. WAO: Let $((l_1,\alpha_1),..,(l_m,\alpha_m)\in\bar{S}^m$ be a vector of linguistic 2-tuples, and $w=(w_1,...,w_m)\in[0,1]^m$ be a weighting vector such that $\sum_{i=1}^{m} w_i=1$. The 2-tuple WAO associated with w is the function $G^w:\langle\bar{S}\rangle^m\to\langle\bar{S}\rangle$ defined by

$$G^w[(l_1,\alpha_1),..,(l_m,\alpha_m)]=\Delta_{\bar{s}}\left(\sum_{i=1}^{m} w_i\beta_i\right) \tag{5}$$

In the same way that experts have importance, criteria also may have it. In this sense we use the process of obtaining the importance of criteria based on the potencies method. This method takes in account the importance for each criterion in the problem solution using a vector of importance with defined values for every criterion involved. When working with linguistic information we just don't have a method for comparing criteria in order to obtain this vector of importance. According to this, it is necessary to obtain the comparison matrix between criteria and then calculate the weighted vector based on criteria importance. The matrix $[A]_{nxn}$ that represents the matrix comparison between criteria is obtained from the experts judgments about criteria. Then the weighted vector ω that represents the weight held by each criterion in the decision process and is obtained using $[A]_{nxn}$ as explained in [14].

Definition 7. WAOC: Let $((l_1,\alpha_1),..,(l_n,\alpha_n)\in\bar{S}^n$ be a vector of linguistic 2-tuples, and $\omega=(\omega_1,...,\omega_n)\in[0,1]^n$ be a weighting vector based on the criteria importance such that $\sum_{j=1}^{n}\omega_j=1$. The 2-tuple aggregation operator associated with ω is the function $G^\omega:\langle\bar{S}\rangle^n\to\langle\bar{S}\rangle$ defined by:

$$G^\omega[(l_1,\alpha_1),..,(l_n,\alpha_n)]=\Delta_{\bar{s}}\left(\sum_{j=1}^{n} w_j\beta_j\right) \tag{6}$$

3 Decision analysis process

Linguistic decision analysis process consists of several phases described below:

Phase 1. Data definition: It defines the evaluation context in which experts will express their preferences. Linguistic descriptors and their semantics are chosen as well as each problem potential solution (alternative) is identified. It also determines the criteria to evaluate every alternative and the experts who are involved in decision

process. In order to allow different expression domain for multiple experts, linguistic terms sets used are organized into an ELH. Therefore, let consider:

A finite set of alternatives $X = \{x_k, k = 1,...,q\}$.

A finite set of criteria $C = \{c_j, j = 1,...,n\}$.

A finite set of experts $E = \{e_i, i = 1,...,m\}$ that express their assessments by using different linguistic scales of information in ELH.

Phase 2. Information gathering: Experts provide their linguistic assessments in utility vectors for each criterion of the evaluated alternatives. The experts express their assessments on every criterion considering every alternative using their linguistic term set in ELH. Due to the fact that our Framework will use linguistic 2-tuple computing model the linguistic preferences provided by the experts will be transformed into linguistic 2-tuples according to the Remark 1.

Phase 3. Computational process: This phase consists of three steps to obtain a global value for each alternative:

-Unification of MGLI. Due to experts provide their assessments in different linguistic scales; it is necessary to transform each assessment in a unique expression domain so called t^* whose granularity is given by Eq. (2). Thus, transformation must be the last level of the ELH according to Eq. (3). Once the information has been unified, it will be expressed by means of linguistic 2-tuples in $S^{n(t^*)}$.

In order to obtain the global value for each alternative the information must be aggregated. In our framework we use four different aggregation operators and the process is performed in two levels:

- Expert Aggregation Level: The first aggregation step it obtains a collective value for all experts' assessments. Here is possible to choose between GMAO and WAO.

- Criteria Aggregation Level: The second one computes a global value for each alternative from results obtained in previous step. Here is possible to choose between AMAO and WAOC. Figure 1 shows the possible combinations of operators.

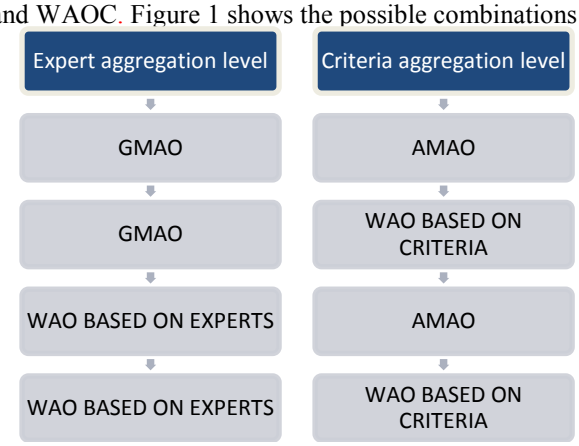


Fig. 1. Framework aggregation operators

Phase 4. Results presentation: Final values are presented in an ordered scale as a ranking of preferences from the most suitable to the less convenient alternative.

4 Illustrative example

We consider the decision process to acquire software in one organization. The decision from where get software implies decide the supply channel option. There are advantages and drawbacks for particular acquire channels and experts many times do not reach to an agreement. In order to satisfy this need, the CEO has arranged a meeting with the three main experts in software solution in the organization: CIO, Head of development department and Head of data management department. The objective of this meeting is to determine which one of the three channels available for software supplying is the most suitable for the company. There are three main channels to obtain software: internal development, external development and buy a standard packet.

In the Internal development, the organization IT department builds the needed software solution.

The External development means acquire by external software development consulting.

Buy a standard package. One of the fastest way for satisfying software needs is by acquiring a standard software package of general purpose. To obtain software 4 criteria should be evaluated, how well it meets the necessary requirements, ease of changes and growths and development time.

Therefore, in phase 1 we have the following:

$$\text{A set of experts } = \begin{Bmatrix} E_1 = \text{CIO}, E_2 = \text{Head of development department}, \\ E_3 = \text{Head of data management department} \end{Bmatrix}$$

$$\text{A set of alternatives} = \begin{Bmatrix} A_1 = \text{Internal development}, A_2 = \text{External development}, \\ A_3 = \text{Buy a standard package} \end{Bmatrix}$$

$$\text{A set of criteria} = \begin{Bmatrix} C_1 = \text{Satisfied requirements}, C_2 = \text{Facility implementing changes}, \\ C_3 = \text{Development time} \end{Bmatrix}$$

An ELH with two linguistic term sets:

$$S_1 = \{VB = \text{Very Bad}, B = \text{Bad}, M = \text{Medium}, G = \text{Good}, VG = \text{Very Good}\}$$

$$S_2 = \begin{Bmatrix} W = \text{Worst}, VB = \text{Very Bad}, B = \text{Bad}, M = \text{Medium}, G = \text{Good}, \\ VG = \text{Very Good}, E = \text{Excellent} \end{Bmatrix}$$

Besides a new level, t^*, in accordance with Eq. (2).

Table 1. Phase 2. Information gathering

Experts	Assessments								
	A_1			A_2			A_3		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
E_1	(E,0)	(VG,0)	(B,0)	(VG,0)	(M,0)	(VG,0)	(M,0)	(W,0)	(E,0)
E_2	(VG,0)	(VG,0)	(VB,0)	(E,0)	(G,0)	(M,0)	(M,0)	(B,0)	(VG,0)
E_3	(VG,0)	(G,0)	(VB,0)	(G,0)	(M,0)	(M,0)	(M,0)	(VB,0)	(VG,0)

Table 2. Criteria comparison

	E_1			E_2			E_3		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
C_1	1	5	1 / 3	1	7	5	1	7	4
C_2	1 / 5	1	1 / 9	1 / 7	1	1 / 2	1 / 7	1	1 / 2
C_3	3	9	1	1 / 5	2	1	1 / 4	2	1

Table 3. Criteria weight vector

ω vector	
Criterion	Weight
C_1	0.2654
C_2	0.0629
C_3	0.6716

Table 4. Experts weight vector

w vector	
Expert	Weight
E_1	0.5
E_2	0.3
E_3	0.2

Table 5. GMAO and AMAO results

Alternative	Percentage
A_2	38,54%
A_1	33,67%
A_3	27,79%

Table 6. WAO with experts weighting and AMAO results

Alternative	Percentage
A_2	37,12%
A_1	33,73%
A_3	29,15%

Table 7. GMAO and WAO based on criteria importance

Alternative	Percentage
A_3	44,34%
A_2	38,25%
A_1	17,42%

Table 8. WAO with experts weighting and WAO based on criteria importance

Alternative	Percentage
A_3	42,62%
A_2	35,98%
A_1	21,41%

Bearing in mind the first step of aggregation, our framework allows use GMAO and WAO in accordance with the weighting vector showed in Table 4. Then, the second aggregation steps we use AMAO and WAO based on criteria importance (see Table 3).

From Table 5 to Table 8 results are expressed in percentage way to better understanding. When it compute GMAO and AMAO (see Table 5) the results are similar to the Table 6 that uses WAO with experts weighting and AMAO (see Table 6). However, a slight difference it can be seen between both but the order of importance is the same. Weighted operator (WAO) introduces a new parameter, the weight of importance of experts, allowing greater differentiation between the final results to elimi-

nate equal importance between opinions. Thus, decision makers have more accurate values with better differentiation between them.

On the contrary, in Tables 7 and 8, the operator for the second level of aggregation used was the weighted vector based on criteria importance. Here, the priority ranking changes significantly. It is because importance vector modifies last criteria aggregation step, allocating highest values for the most important criterion and reducing values for the others. Furthermore, obtained values in Table 8 take into account the weight of the experts.

5 Conclusion

Aggregation refers to the process of combining several values into a single one, so that the final result of aggregation takes into account in a given manner all the individual values. Such an operation is used in many well-known disciplines such as Multi-Criteria Multi-Expert Decision Making. In order to reach good results for decision process, classical synthesizing functions have been proposed: arithmetic mean, geometric mean, median and many others. In this papers we present a linguistic framework developed that allows analyze different decision results by using several aggregation operators. In this regard we also propose compute criteria importance based on the potencies method with Saaty scale.

Currently, the framework computation capability is expanded by using different aggregation operators such as Ordered Weighted Averaging (OWA) aggregation operators' family. In addition, we are comparing different methodologies and decision making approaches such as Analytic Hierarchy Process (AHP).

References

1. Figueira, J., Greco, S., Ehrgott, M.: Multiple Criteria Decision Analysis: State of the Art Surveys. Kluwer Academic Publishers, Boston/Dordrecht/London (2005)
2. Clemen, R.: Making Hard Decisions. An Introduction to Decision Analisys. Duxbury Press (1995)
3. Zadeh, L.: The concept of a linguistic variable and its applications to approximate reasoning. Information Sciences, Part I, II, III (8,9) (1975) 199–249,301–357,43–80
4. Dong, Y., Xu, Y., Yu, S.: Linguistic multiperson decision making based on the use of multiple preference relations. Fuzzy Sets and Systems 160(5) (2009) 603–623
5. Delgado, M., Verdegay, J., Vila, M.: Linguistic decision-making models. International Journal of Intelligent Systems 7(5) (1992) 479–492
6. Herrera, F., Herrera-Viedma, E., Martínez, L.: A fusion approach for managing multi-granularity linguistic term sets in decision making. Fuzzy Sets and Systems 114(1) (2000) 43–58
7. Herrera, F., Herrera-Viedma, E., Verdegay, J.: A linguistic decision process in group decision making. Group Decision and Negotiation 5 (1996) 165–176
8. Martínez, L., Herrera, F.: An overview on the 2-tuple linguistic model for computing with words in decision making: Extensions, applications and challenges. Information Sciences 207 (2012) 1–18

9. Degani, R., Bortolan, G.: The problem of linguistic approximation in clinical decision making. International Journal of Approximate Reasoning 2 (1988) 143–162
10. Delgado, M., Verdegay, J., Vila, M.: On aggregation operations of linguistic labels. International Journal of Intelligent Systems 8(3) (1993) 351–370
11. Herrera, F., Martínez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems 8(6) (2000) 746–752
12. Herrera, F., Martínez, L.: The 2-tuple linguistic computational model. Advantages of its linguistic description, accuracy and consistency. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9 (2001) 33–48
13. Espinilla, M., Liu, J., Martínez, L.: An extended hierarchical linguistic model for decision-making problems. Computational Intelligence 27(3) (2011) 489–512
14. Saaty, R. W.: The analytic hierarchy process: What it is and how it is used. Math Modelling, Vol. 9, No. 3-5, pp. 161-176 (1987).

IV WORKSHOP PROCESAMIENTO DE SEÑALES Y SISTEMAS DE TIEMPO REAL

- WPSTR -

CACIC2013

IV WORKSHOP PROCESAMIENTO DE SEÑALES Y SISTEMAS DE TIEMPO REAL

- WPSTR -

ID	Trabajo	Autores
5708	**Determinación de modelos para el análisis del control y estabilidad de sistemas dinámicos**	Carlos Alvarez Picaza (UNNE), María Ines Pisarello (UNNE), Jorge E. Monzón (UNNE)
5806	**Framework para modelado de Transacciones en Sistemas de Bases de Datos de Tiempo Real**	Carlos Buckle (UNPSJB), José M. Urriza (UNPSJB), Damián Pablo Barry (UNPSJB), Lucas Schorb (UNPSJB)
5607	**Toolchain and workflow for the design of an ISO 11783-compatible ECU based on ISOAgLib**	Joaquin Ezpeleta (UNR), Sebastian Rossi (UNR)
5842	**Controlador neuronal incremental aplicado a un mezclador de flujos**	Sergio L. Martínez (UNJu), Enrique E. Tarifa (UNJu), Samuel Franco Dominguez (UNJu)
5608	**A Fault Resilience Tool for Embedded Real-Time Systems**	Franklin Lima Santos (UFB), Flavia Maristela Santos Nascimento (IFBa)
5860	**Application of Zigbee Technology for Monitoring Environmental Variables in Greenhouses**	Juan Carlos Suárez Barón (UNAD)
5675	**Inversión de prioridades: prueba de concepto y análisis de soluciones**	Raúl Benencia (UNLP), Fernando Romero (UNLP), Fernando Tinetti (UNLP), Luciano Iglesias (UNLP)
5677	**Estudio sobre mediciones de Campos Electromagnéticos No Ionizantes**	Jorge S. García Guibout (UDA), Miguel Mendez Garabetti (UDA), Antonio Castro Letchtaler (IESE), Alfredo David Priori (UA)

IV WORKSHOP PROCESAMIENTO DE SEÑALES Y SISTEMAS DE TIEMPO REAL

- WPSTR -

ID	Trabajo	Autores
5729	**Selección sub-óptima del espectro asociado a la matriz de afinidad**	Luciano Lorenti (UNLP), Lucía Violini (UNLP), Javier Giacomantone (UNLP)
5686	**Isolated Spanish Digit Recognition based on Audio-Visual Features**	Gonzalo Sad (UNR), Lucas Terissi (UNR), Juan Carlos Gómez (UNR)
5820	**Desarrollo de una Ficha Anestésica Web en Áreas críticas**	Gustavo Bianco (HIBA), Marcelo Sabalza (HIBA),Daniel Luna (HIBA), Gustavo Garcia Fornari (HIBA), Jorge Garbino (HIBA),Martin Waldhorn (HIBA), Estefanía Tarsetti (HIBA)
5856	**Detección de signos respiratorios patológicos en poblaciones avícolas productivas mediante procesamiento digital de señales acústicas**	César E. Martínez (UNL), Cristian Kühn (UNL)
5649	**Sievert-type measurement and acquisition system for the study of hydrogen storage in solids**	Jorge Runco (UNLP),Marcos Meyer (UNLP)

Determinación de modelos para el análisis del control y estabilidad de sistemas dinámicos

Carlos Álvarez Picaza[1], María Inés Pisarello[1], Jorge E. Monzón[1]

[1] Dpto de Ingeniería. Fac de Ciencias Exactas. Universidad Nacional del Nordeste
Corrientes, Argentina
cpicaza@gmail.com

Abstract. Los modelos aquí presentados analizan las condiciones de control y estabilidad de dos sistemas dinámicos de distinta naturaleza, uno eléctrico y otro biomédico. Los modelos son desarrollados en el espacio de estados, lo que aporta nuevas respuestas al análisis de sistemas.

Keywords: espacio de estados, controlabilidad, estabilidad.

1 Introducción

La Teoría de Control Clásico describe al sistema dinámico a través de la relación matemática entre la entrada y la salida, o sea su función transferencia, considerando en general a este sistema dinámico como una "caja negra". Esto se muestra en la **Fig 1**, en la cual se observa que a través de la inyección de diferentes tipos de señales a la entrada de la caja negra se obtiene un conjunto de señales a la salida de la misma, lo que nos permite conocer el comportamiento del sistema dinámico y así definir las propiedades de este sistema [1]. Entonces a partir de estos ensayos es posible establecer una relación matemática de la función de transferencia del sistema en cuestión, dado por:

$$g(t) = \frac{y(t)}{u(t)} \tag{1}$$

Fig. 1. Modelo general de caja negra

Normalmente los análisis de datos se realizan desde el punto de vista gráfico en el espacio de los tiempos y en el espacio de las frecuencias, mediante la utilización de la función transferencia correspondiente. Lo que permite la Teoría de Control Moderno es el análisis en otro contexto gráfico conocido como "espacio de estados", a partir del cual, se puede inferir nueva información [1,2].

Espacio de Estado: Es el espacio n-dimensional cuyos ejes de coordenadas están formados por el eje x_1, eje x_2, … , eje x_n, donde x_1, x_2, … ,x_n son las variables de estado. Cualquier estado se puede representar como un punto en el espacio de estados.

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad \textit{Ec. de Estado}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \quad \textit{Ec. de Salida}$$

donde A se denomina matriz de estado, B matriz de entrada, C matriz de salida y D matriz de transmisión directa.

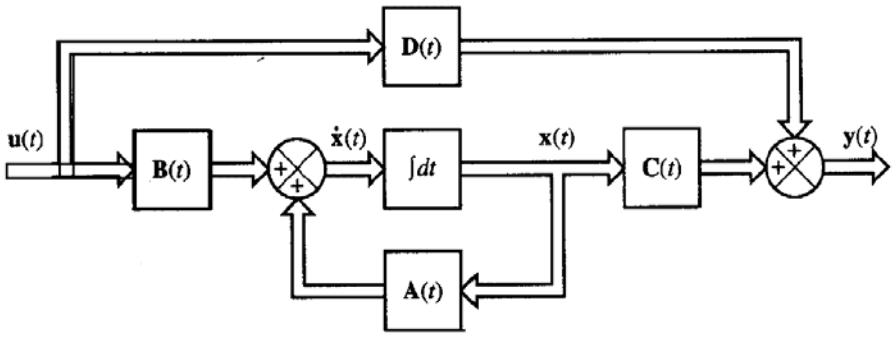


Fig. 2. Diagrama en bloque del sistema de control en el espacio de estados.

El objetivo de este trabajo es determinar el comportamiento de dos sistemas de distinta naturaleza, uno biomédico (sistema cardiovascular) y otro eléctrico-electrónico (turbina eólica) mediante el análisis de sus funciones de transferencia, es decir sus señales de entrada y salida. Para ello diseñamos un modelo de tipo caja negra en el espacio de estados.

2 Métodos

Utilizaremos el modelado en el espacio de estados aplicado a señales de una turbina eólica (energías renovables) y de un electrocardiograma (biomédica).

La señal de entrada para un sistema de control no se conoce con anticipación, pero es de naturaleza aleatoria, y la entrada instantánea no puede expresarse en forma analítica. Solo en algunos casos especiales se conoce con anticipación la señal de entrada y se puede expresar en forma analítica o mediante curvas [3].

En el análisis y diseño de sistemas de control, debemos tener una base de comparación del desempeño de algunos. Esta base se configura especificando las señales de entrada de prueba particulares y comparando las respuestas de varios sistemas a estas señales de entrada.

Muchos criterios de diseño se basan en tales señales o en la respuesta del sistema a los cambios en las condiciones iniciales (sin señales de prueba). El uso de señales de prueba se justifica porque existe una correlación entre las características de respuesta de un sistema para una señal de entrada de prueba común y la capacidad del sistema de manejar las señales de entrada reales.

Señales de prueba típicas. Las señales de prueba que se usan regularmente son funciones escalón, rampa, parábola, impulso, senoidales, etc. Con estas señales de prueba, es posible realizar con facilidad análisis matemáticos y experimentales de sistemas de control, dado que las señales son funciones del tiempo muy simples [3].

2.1 Aerogenerador:

Debido a la complejidad del modelo del generador de inducción, el principio del proceso de auto-excitación se explica con el uso de un circuito RLC debido a que el comportamiento del generador de inducción es similar a un circuito de este tipo [4].

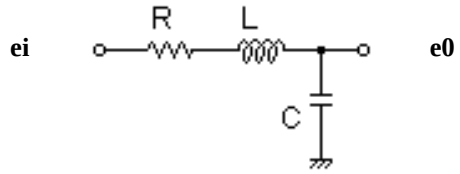


Fig. 3. Modelo para representación en el espacio de estados del generador

Donde R representa la resistencia equivalente rotórica y estatórica, L los devanados de estator y C, un banco de capacitares reemplazando a la tensión de línea.

Para el modelo objeto de tratamiento la función transferencia es

$$G(s) = \frac{\frac{1}{LC}}{s^2 + \frac{R}{L}s + \frac{1}{LC}} = \frac{E_0(s)}{E_i(s)} \qquad (2)$$

Para definir las variables de estado

$$\ddot{e}_o + \frac{R}{L}.\dot{e}_o + \frac{1}{LC}.e_o = \frac{1}{LC}.e_i \tag{3}$$

$$x_1 = e_o \tag{4}$$

$$x_2 = \dot{e}_o \tag{5}$$

y las variables de entrada y salida mediante

$$u = e_i \tag{6}$$

$$y = e_o = x_1 \tag{7}$$

Matricialmente, queda planteada la siguiente ecuación de estado

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{-1}{LC} & \frac{-R}{L} \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{LC} \end{bmatrix} .u \tag{8}$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{9}$$

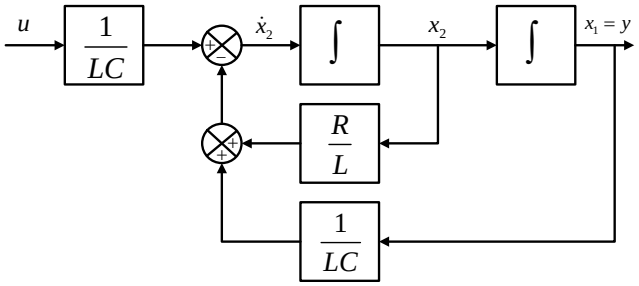


Fig. 4. Diagrama en bloques funcional del aerogenerador representado en el espacio de estados

2.2 Pared Cardíaca:

En este trabajo modelamos la dinámica cardíaca utilizando un modelo de Windkessel de tres elementos. El modelo fue elaborado en el espacio de estados. Este enfoque nos permite obtener resultados acerca de la estabilidad del sistema [5].

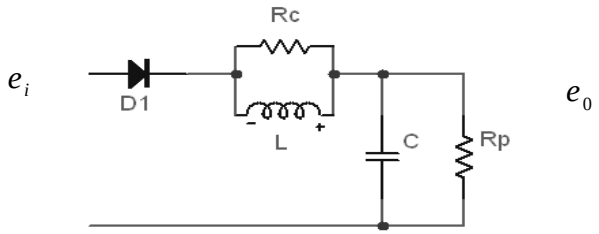


Fig. 6. - Modelo Windkessel de 3 elementos

El modelo consiste en una conexión paralela de una resistencia y un capacitor. La resistencia *Rp* representa la resistencia total periférica y el capacitor *C* representa la compliancia de los vasos. Otro elemento resistivo entre la bomba y la cámara de aire, *Rc*, simula la resistencia del flujo sanguíneo debido a la válvula aórtica o pulmonar. *L* es un elemento inercial en paralelo con la resistencia característica, *Rc*. Con estos arreglos, el modelo cuenta con la inercia de todo el sistema arterial a bajas frecuencias y a altas y medias frecuencias permiten que intervenga la resistencia característica [6].

Generalizando $R_p = R$ y R_c muy pequeña, para el modelo objeto de tratamiento es

$$G(s) = \frac{\frac{1}{LC}}{s^2 + \frac{1}{RC}s + \frac{1}{LC}} = \frac{E_0(s)}{E_i(s)} \tag{10}$$

obteniendo

$$s^2.E_0(s) + s.E_0(s).\frac{1}{RC} + E_0(s).\frac{1}{LC} = E_i(s).\frac{1}{LC} \tag{11}$$

para definir las variables de estado

$$\ddot{e}_o + \frac{1}{RC}.\dot{e}_o + \frac{1}{LC}.e_o = \frac{1}{LC}.e_i \tag{12}$$

Quedando planteada la siguiente ecuación de estado

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{-1}{LC} & \frac{-1}{RC} \end{bmatrix}.\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{LC} \end{bmatrix}.u \tag{13}$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix}.\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{14}$$

Teniendo en cuenta estos conceptos, el objetivo de este trabajo es aplicar las ecuaciones de estado a los distintos sistemas para el modelado correspondiente y posterior análisis de controlabilidad.

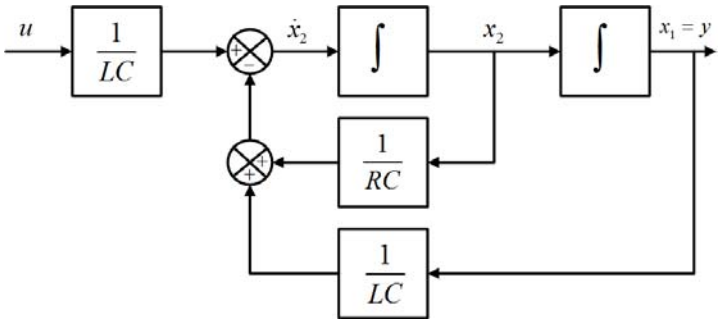


Fig. 7. Diagrama en bloques funcional de la pared cardíaca representada en el espacio de estados

3. Resultados y Discusión

31. Aerogenerador

De acuerdo a datos de ensayos en cortocircuito realizado a un generador de 5 KVA y considerando el condensador para estabilizar a la salida, se obtiene

Tabla 1. Características eléctricas del sistema generador

X_L (ohm)	R (ohm)	C (*u*F)	Sigma
1,5	0,55	90	1

Donde las curvas corresponden a la variación de x_1 y la variación de x_2 (variables de estado).

3.2 Pared Cardíaca

Tabla 2. Características funcionales del sistema pared cardíaca

Señal	PS (mmHg)	PD (mm Hg)	L (PS-PD)	C (e-4cm/mm HG)
a41770	93,28	57,12	36,16	7,09

Las curvas correspondientes a estos resultados están definidas en las **Fig 8** a **11**.

3.3 Controlabilidad y Estabilidad

La controlabilidad es una de las propiedades cualitativas de los sistemas dinámicos. A grandes rasgos, la controlabilidad estudia la posibilidad de guiar o llevar los estados de un sistema hacia una posición deseada mediante la señal de entrada [6].

Dada una matriz de Controlabilidad genérica

$$\mathscr{C} = \left[\mathbf{B}\ \mathbf{AB}\ \mathbf{A}^2\mathbf{B}\ \ldots\ \mathbf{A}^{n-1}\mathbf{B}\right] \tag{15}$$

Si el rango de $\mathscr{C}$ = $n \rightarrow$ Existe una entrada que hace que el sistema pase de cualquier estado inicial al estado final deseado.

Señal Aerogenerador:

$$\mathscr{C} = \begin{bmatrix} 0 & 7.4074 \\ 7.4074 & -2.7163 \end{bmatrix}$$

$$\Delta = -54.8697\ ,\ \text{rango} = 2$$

$$\lambda_{\min} = -8.889$$

$$\lambda_{\max} = 6.1728$$

$$cond(\mathscr{C}) \approx -0.6944$$

$$G'(s) = \frac{0.6944}{s^2 + 0.3667s + 0.6944}$$

La **Fig. 8** muestra el efecto de la condición de control, aplicada al sistema, indicando una disminución del valor de la frecuencia natural del sistema para nuevos valores de entrada (x1c y x2c).

De acuerdo a los polos de la nueva función transferencia, el sistema se hace mas amortiguado, lo que conlleva a una mas rápida estabilidad al nivel 0.

Señal Pared Cardíaca:

$$\mathscr{C} = \begin{bmatrix} 0 & 0.0039 \\ 0.0039 & -0.0005 \end{bmatrix}$$

$$\Delta = -1.5214*10^5\ ,\ \text{rango} = 2$$

$$\lambda_{\min} = -0.0042$$

$$\lambda_{\max} = 0.0036$$

$$cond(\mathscr{C}) \approx -0.8571$$

$$G'(s) = \frac{0.8571}{s^2 + 0.1370s + 0.8571}$$

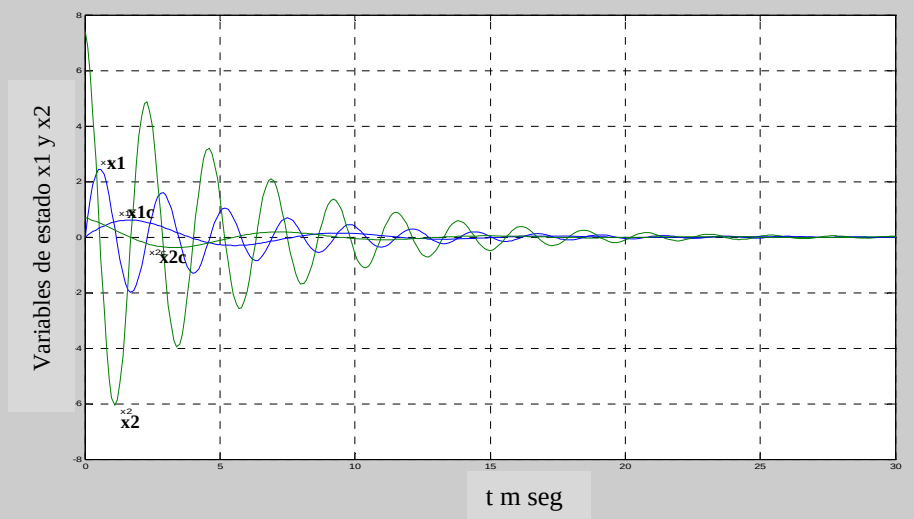


Fig. 8 –Respuesta natural del sistema generador incluyendo la condición de control

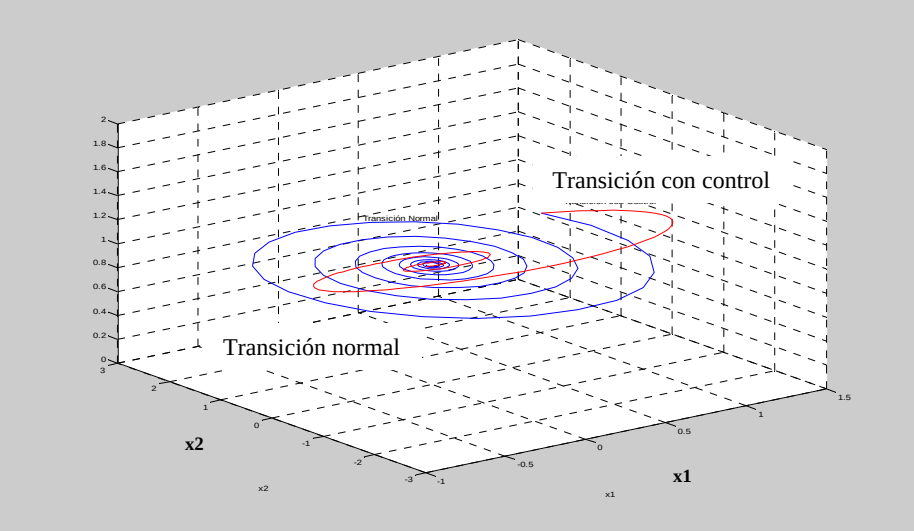


Fig. 9. – Espacio de estados - Sistema Aerogenerador. 3D

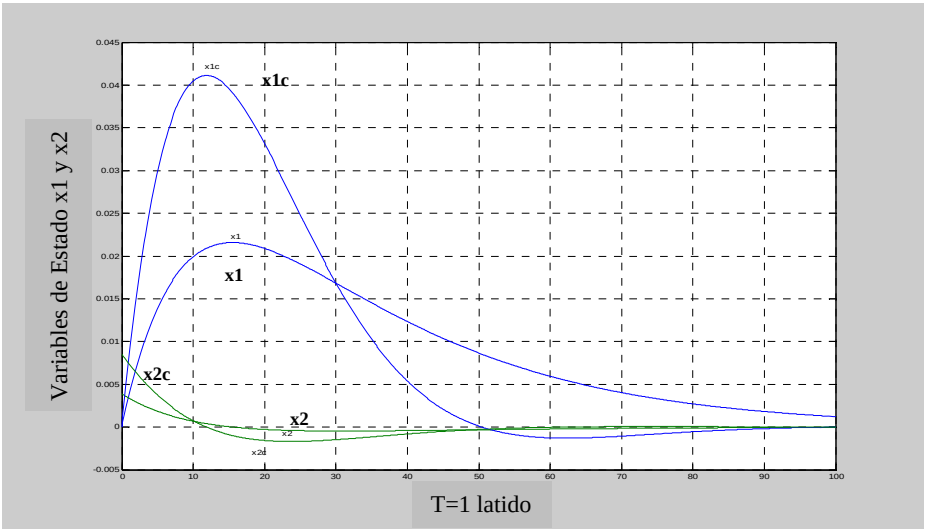


Fig. 10. Respuesta natural del sistema generador incluyendo la condición de control

La incorporación de la condición de control, produce en el sistema de estudio un aumento de la frecuencia natural del mismo, haciendo que el sistema alcance su equilibrio en un tiempo menor (hay que tener en cuenta que la frecuencia natural del sistema se encuentra acotada a 1 latido/seg.).

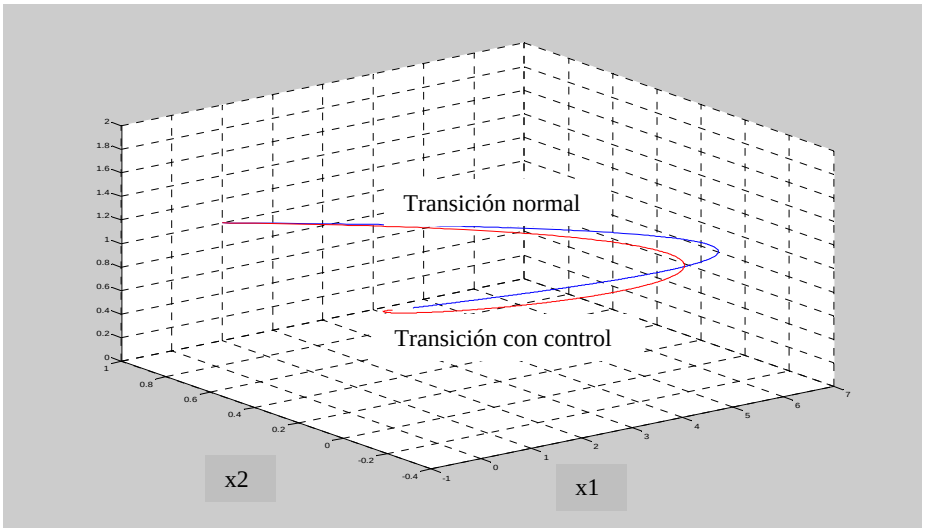


Fig. 11. Espacio de estados - sistema pared cardíaca 3D

4. Conclusiones

El hecho de que al utilizar la condición de control, ambos sistemas acortan notablemente la trayectoria desde un estado genérico $x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ hacia el punto estable $x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, es decir, al utilizar estas herramientas de control el sistema gana estabilidad haciendo a "cualquier sistema" mas controlable y asegurando su convergencia a su estado de equilibrio.

Referencias

1. Ogata K.: Ingeniería de control Moderna. 4ta Edición Ed Pearson. ISBN: 84-205-3678-4.
2. Kuo B.: Sistemas de Control Automático. 7ma Edición Ed. Prentice Hall Hispanoamericana S.A. ISBN: 968-880-723-0.
3. Rautenberg C., D'Attellis C.: Control lineal Avanzado y Control Óptimo. 1era Edición Ed. Argentina-AADECA. ISBN: 987-20960-5253.
4. Henández Sanchez A.M.: Análisis, Modelado y Simulación de la Operación de Sistemas de Generación Eoleoeléctrica Basados en Generadores de Inducción de Tipo Jaula de Ardilla. Centro Nacional de Investigación y desarrollo Tecnológico. Méjico. 2008.
5. Monzón J.E., Álvarez Picaza C., Pisarello M.I.: A multiple-input multiple-output system for modeling the cardiac dinamycs. In: Proceedings of the 33th Annual International Conference of the IEEE-EMBS. 2011
6. Álvarez Picaza C., Pisarello M.I., Monzón J.E.: Modelado del Sistema Cardíaco en el Espacio de Estados aplicando Criterios de Controlabilidad. In: 1er Congreso de Bioingeniería Costa Rica 2009. San José de Costa Rica. 2009

Framework para modelado de Transacciones en Sistemas de Bases de Datos de Tiempo Real

Carlos E. Buckle, José M. Urriza, Damián P. Barry, Lucas Schorb

Depto Informática, Facultad Ingeniería, Universidad Nacional de La Patagonia San Juan Bosco
Puerto Madryn, Argentina
+54 280-4472885 – Int. 117.
cbuckle@unpata.edu.ar, josemurriza@unp.edu.ar, damian_barry@unpata.edu.ar
.

Resumen. En los Sistemas de Bases de Datos de Tiempo Real, se requiere extender el modelo tradicional de transacciones para incorporar restricciones temporales. En este tipo de sistemas, una transacción se debe ejecutar dentro de un intervalo de tiempo específico, predeterminado con anterioridad. Además, se debe garantizar la validez de ciertos datos, cuyo valor caduca con el paso del tiempo. Los conceptos teóricos para modelado de transacciones de tiempo real, han sido desarrollados en varios trabajos del área. En este trabajo, se reúnen y exponen los conceptos principales en lo referido a consistencia temporal de datos y transacciones. Luego, se los incorpora a un framework orientado a objetos que guía y facilita el modelado de transacciones con restricciones de tiempo. El uso del framework se presenta y se verifica aplicándolo sobre un caso de ejemplo.

Palabras Clave: Transacciones de Tiempo Real, Consistencia Temporal, Bases de Datos de Tiempo Real, Ingeniería de Software de Tiempo Real, Datos de Tiempo Real.

1 Introducción

El procesamiento de *transacciones* ([1]), es la técnica utilizada para el acceso concurrente, consistente y tolerante a fallas en un Sistema de Bases de Datos (*DBS*) convencional. Los Sistemas de Tiempo Real (*RTS*) que manejan datos persistentes sobre un *DBS,* imponen que el modelo de transacciones sea ampliado para considerar restricciones temporales. Por ejemplo, en sistemas de supervisión y monitoreo de procesos (*SCADA*), en tableros de comando (*Balanced Scorecard*), en ambientes de transacciones on-line de compra/venta de acciones bursátiles (*Online Stock Trading*), etc. Esta aproximación entre las disciplinas de *DBS* y *RTS,* ha dado origen a los *Sistemas de Bases de Datos de Tiempo Real* (*RTDBS*) ([2, 3]). Los cuales, además de administrar un gran volumen de datos convencionales, deben manejar datos que se encargan de reflejar el estado de elementos variables del ambiente. Estos datos, tienen como característica que su valor *envejece,* hasta perder vigencia una vez alcanzado su

vencimiento (*Datadeadline* [4]). Por ese motivo, son identificados en el sistema como *datos de tiempo real* (*RTD Real-Time Data*). Mantener correctamente actualizados estos objetos, permite garantizar la consistencia entre el ambiente supervisado y el *RTDBS*. En este escenario, las transacciones además de garantizar consistencia lógica, deben garantizar *consistencia temporal* ([2]), dando origen a lo que se denomina *transacciones de tiempo real* (*RTT*) (*Real-Time Transaction*) ([5], [6]). Sobre ellas debe garantizarse: su instante mínimo de inicio (*arrival time*), su tiempo de respuesta previo al vencimiento (*deadline*) y la validez de los *RTD* involucrados.

Estas consideraciones, introducen una complejidad adicional al momento de construir aplicaciones. Además de la lógica propia del dominio del problema a resolver, deben atenderse otras cuestiones como: la contabilización de tiempos, la validación de reglas y la planificación de tareas para garantizar la *consistencia temporal* de datos y transacciones. Surge así la necesidad de contar con herramientas que incorporen estos conceptos, orienten el modelado y aporten la información temporal necesaria para que el planificador del *RTS* pueda ordenar debidamente la ejecución de las transacciones.

En el presente trabajo se presenta un *framework* orientado a objetos, para el modelado de *RTT*, como un aporte concreto para los diseñadores de *RTDBS*.

Este documento está organizado de la siguiente manera: en la sección 2 se presentan trabajos anteriores relacionados con el modelado de *RTDS*. En la sección 3 se presenta un conjunto de definiciones y clasificaciones asociadas con el concepto de *RTT*. En la sección 4 se presenta el *framework* desarrollado. En la sección 5 se aplica el *framework* sobre un caso de estudio. En la sección 6 se elaboran las conclusiones y se plantean los trabajos a futuro.

2 Trabajos previos

En el pasado, se han desarrollado una serie de trabajos que apuntan a modelar objetos y transacciones dentro de un *RTDBS*. Uno de los trabajos más importantes en el modelado orientado a objetos es *Real Time Semantic Objects Relationships And Constraints* (*RTSORAC*) ([7]), en el cual se definen tres propiedades básicas de los *RTDBS*: objetos, relaciones y transacciones de tiempo real. El modelo teórico contempla la mayoría de los requerimientos temporales y permite expresar las diferentes entidades que intervienen en un *RTDBS*. Sin embargo, es demasiado extenso y genérico como para ser aplicado directamente en el diseño de aplicaciones concretas. No obstante, es utilizado en un conjunto de trabajos posteriores que lo utilizan como base para la definición de perfiles de diseño, como los presentados en [8] y [9]. Dichos trabajos, se enfocan preferentemente a la definición de objetos de tiempo real y no al modelado de transacciones. Tampoco se consideran objetos de cambio discreto ([10]), ni datos derivados (datos calculados). El patrón de diseño presentado en [11], incorpora el modelado de transacciones pero solo de aquellas encargadas de actualizar datos desde sensores. El marco de trabajo (*framework)* presentado en [12], clasifica aquellas transacciones que actualizan los datos desde

sensores y aquellas que propagan derivaciones de los datos calculados, pero no incluye a las transacciones del usuario que resuelven la lógica de la aplicación.

En lo que sigue, se presenta un *framework* orientado a objetos para el modelado de *RTT*, en el cual se consideran aspectos ampliados respecto de los trabajos antes mencionados. El *framework* presentado utiliza el *Tipo de Dato Abstracto para Bases de Datos de Tiempo Real RTD* ([13]), resultante de un trabajo anterior, el cual se toma como base y permite definir cualquier tipo de atributo de tiempo real encapsulando la validación de *consistencia temporal*.

3 Gestión de Transacciones con Restricciones Temporales

El modelo de *RTT,* puede definirse como una extensión del modelo tradicional de transacciones ([14]), el cual se resume brevemente a continuación: Un *DBS* es un conjunto de *entidades de datos,* que representan información sobre un contexto determinado. El mapeo entre entidades y sus valores, define el *estado* de la base de datos en un determinado instante. Sobre ella se definen un conjunto de operaciones que permiten recuperar, crear, modificar y eliminar entidades. Estas operaciones provocan la transición de un estado a otro. Una *transacción*, es un conjunto de operaciones parcialmente ordenadas sobre la base de datos que debe ser ejecutada atómicamente. El orden parcial de las operaciones está dado por un algoritmo. La *atomicidad* significa que la transacción debe ser ejecutada satisfactoriamente o sino no debe ser ejecutada. Para esto el *DBS* ofrece dos operaciones: *commit,* que confirma la finalización satisfactoria de una transacción y *rollback,* que deshace la ejecución parcial de operaciones realizadas por una transacción que no puede finalizar satisfactoriamente. La transacción pasa de un estado *consistente* de la base de datos, a un próximo estado consistente. La ejecución de las operaciones ordenadas de una transacción se asume correcta, si se ejecuta en forma aislada. El *aislamiento,* oculta los cambios parciales que realiza una transacción hasta su finalización. Si la transacción finaliza con *commit* se garantiza la *durabilidad* (persistencia) de los resultados en la base de datos. De esta forma se han definido las características *ACID* (*atomicidad*, *consistencia*, *aislamiento* y *durabilidad*).

Sobre este modelo tradicional de transacciones, se puede incorporar la noción de *tiempo real*. Una transacción puede conocer el *tiempo actual* (*now*) accediendo a una entidad única del sistema llamada *reloj* (*clock)*. Esta entidad es solo-lectura y toma valores positivos que se incrementan monotónicamente en concordancia con el paso del tiempo. Esto permite establecer intervalos de vigencia sobre los *RTD* y establecer restricciones temporales sobre las transacciones. Si una transacción no cumple estas restricciones temporales al momento del *commit*, se debe deshacer (*rollback*) y si corresponde, volverse a ejecutar (*restart*).

3.1 Consistencia Temporal de los Datos de Tiempo Real

Los *RTD* reflejan objetos cambiantes del ambiente. Existen dos tipos ([13]): Los *RTDBase*, los cuales reflejan el estado de un objeto externo y actualizan su valor desde sensores o publicadores y los *RTDDerived* que son datos derivados, cuyo valor se determina con cálculos sobre un *Conjunto-Lectura* (*Read Set*) con otros *RTD*.

La *consistencia temporal* de los *RTD* garantiza:

- *validez temporal*: El valor de un *RTD* se considera válido dentro de un *intervalo de validez*, *VI* (*validity interval*). En el cual el límite inferior (*VILB*) es el momento en el que se actualiza el valor y el límite superior (*VIUB*) es el instante en el que el valor pierde vigencia (*DataDeadline*). Un *RTD* respeta *consistencia temporal absoluta* ([10]) en el instante *t* si $VILB_{dtr}(t) \leq now(t) \leq VIUB_{dtr}(t)$.
- *coherencia temporal*: Se debe garantizar que los datos calculados se deriven en base a los *RTD* actualizados en instantes cercanos de tiempo. Por esto, para los *RTDDerived* se debe garantizar *consistencia temporal relativa* ([10]) sobre todo su conjunto lectura, es decir, $\bigcap\{VI_x(t) | x \in \text{ReadSet}_{dtr}\} \neq \varnothing$ y su *VI* se calcula como:
  $VILB_{dtr}(t) = Max\{VILB_x(t) | x \in \text{ReadSet}_{dtr}\}$ y $VIUB_{dtr}(t) = Min\{VIUB_x(t) | x \in \text{ReadSet}_{dtr}\}$

Los *RTDBase* pueden ser continuos (*RTDBaseContinuous*) ó discretos (*RTDBaseDiscrete*). Los continuos, son actualizados con muestras periódicas y su *VI* depende de su *edad* (tiempo desde su última actualización). El vencimiento del dato se define en base a establecer la *edad máxima* (*maximumAge*) ([15]) y es posible garantizar la *consistencia temporal absoluta*, si se considera un período de actualización $P \leq maximumAge/2$ ([4]). Estos conceptos se desarrollaron en la definición del tipo de dato *abstracto RTD* presentado en [13].

La actualización de los *RTDBase* puede implementarse utilizando dos políticas: *actualización inmediata* o *a demanda* ([16]). La *actualización inmediata* garantiza que cada cambio en un objeto externo será inmediatamente reflejado en el *RTDBase* que lo representa. La política de actualización *a demanda*, significa que el *RTDBase* será actualizado solo cuando una transacción necesite utilizarlo. La *actualización inmediata* procura un sistema más predecible, pero genera una carga innecesaria al actualizar valores de *RTD* que quizás no requiera ninguna transacción. La *actualización a demanda* soluciona este problema, pero introduce un tiempo de latencia en las transacciones pues deben actualizar los *RTDBase* antes de utilizarlos.

3.2 Características de las Transacciones de Tiempo Real.

Como fue presentado inicialmente, una *RTT* tiene las restricciones temporales propias: un tiempo mínimo de inicio (*startTime*) y un tiempo de respuesta anterior al vencimiento (*deadline*). Además, las *RTT* están condicionadas por la validez temporal de los *RTD* involucrados en ella. En un *RTDBS* se identifican tres clases de *RTT*:

- *RTT de Actualización de RTDBase* (*RTTUpdate*): Son transacciones de solo-escritura sobre un conjunto de *RTDBase,* llamado *Write Set*. El sistema debe implementarlas para garantizar la *consistencia temporal absoluta* de los objetos del *Write Set*.
- *RTT de Derivación de RTDDerived* (*RTTDerivation*): Son transacciones de lectura sobre un *Read Set* de *RTD* y de escritura sobre un *Write Set* de *RTD*. El sistema debe implementarlas para garantizar el re-cálculo de datos derivados.

- *RTT del Usuario* (*RTTUsrApp*): Implementan la lógica de la aplicación de usuario. Son transacciones de solo-lectura sobre un *Read Set* de *RTD*, aunque además, utilizan datos convencionales que no tienen restricciones de tiempo real.

Dependiendo de la política de actualización de los *RTDBase*, las *RTTUpdate* se pueden implementar como transacciones independientes (en *actualización inmediata*), o como sub-transacciones de una *RTTUsrApp* (en *actualización a demanda*). Las transacciones independientes se deben planificar con un período *P* igual al menor período de los *RTDBase* en el *Write Set.* Su *DataDeadline* se puede calcular como el menor *VIUB* de dicho conjunto, generalmente como *P*2* ([4]).

Las *RTTDerived* pueden ser transacciones independientes o transacciones disparadas (*triggered*) por una *RTTUpdate* [16]. Si son independientes se debe planificar con un período *P* igual al menor período de los *RTD* en el *Read Set*. Su *DataDeadline* es el menor *VIUB* de todos los *RTD* incluidos en el *Read Set*.

Cada *RTT,* independientemente de su clase, puede tener su propio requerimiento de tiempo de respuesta, por ende, debe poder indicarse un tiempo para el vencimiento (*timeToDeadline*) propio de la transacción. Si no se explicita un *timeToDeadline* se puede considerar como máximo un *timeToDeadline = P* (período de la *RTT*). Para que sea posible la planificación, también es necesario poder estimar el peor caso de tiempo de ejecución de cada transacción (*Worst Case Execution Time – WCET*).

3.3 Planificación de Transacciones

Para garantizar las restricciones temporales mencionadas, es necesario que el *RTDBS* cuente con un planificador de tiempo real (*real-time scheduler*) que ordene la ejecución concurrente de transacciones. Para que esto sea posible, se debe manejar un conjunto de atributos y medidores de performance de cada *RTT*. Mínimamente se requiere:

- *arrivalTime*: Instante en que la *RTT* se pone lista para ejecutar.
- *startTime*: Instante en que la *RTT* comienza a ejecutar.
- *WCET*. *Worst case execution time:* Tiempo máximo de ejecución estimado.
- *period*: Magnitud del período, para aquellas *RTT* de ejecución periódica.
- *timeToDeadline*: lapso de tiempo para el *deadline.*
- *deadline*: vencimiento de la *RTT= (arrivalTime + timeToDeadline).*
- *dataDeadline(t)*: Menor *DataDeadline* de los *RTD* del *Read Set* al instante *t*.
- *estRemaining(t)*: lapso de ejecución remanente de la *RTT* al instante *t.*
- *estCompletionTime(t)*: Instante estimado de fin en *t* = (*t* + *estRemaining(t)*).
- *estSlack(t)*: lapso que se puede retrasar = *deadline – estCompletionTime(t)*.

En función de algunos de estos parámetros, el planificador de tiempo real debe determinar en cada instante *t* de planificación la prioridad de la *RTT*.

- $priority_{rtt}(t)$: prioridad de la *RTT* en el instante *t*.

El algoritmo de planificación, debe garantizar que las *RTT* del sistema se ejecuten antes de su *deadline* y que cumplan con las reglas de *consistencia temporal*. Para esto se debe implementar una política de planificación. Algunas de las posibles son:

- *Rate Monotonic* (*RM*) ([17]): Prioridad basada en la magnitud del *período*.
- *Earliest-Deadline-First* (*EDF*) ([17]): Prioridad basada en el *deadline.*
- *Earliest-DataDeadline-First* (*EDDF*) ([4]): Prioridad basada en el *dataDeadline.*
- *Highest-Value-First (HVF)* ([18]): Prioridad basada en el la *función valor* (Fig. 1).
- *Least-Slack-First* (*LSF*)([5]): Prioridad basada en el tiempo disponible (*estSlack*).
- *Shortest-Job-First (SJF):* Prioridad basada en el tiempo que resta (*estRemaining*).

Estas políticas se pueden mejorar con el agregado de los conceptos de *función valor* y *espera inducida*. Estos se describen en los siguientes párrafos.

Las *RTT* pueden tener diferentes niveles de criticidad en su vencimiento. Estos niveles de criticidad, influyen en las decisiones del planificador, el cual puede considerar el concepto de *Función Valor* introducido por Jensen en [18]. La *función valor* de una transacción permite medir el *valor* (*ganancia*) del sistema si finaliza la transacción en un determinado instante *t*. Los valores positivos son deseables y los negativos indeseables. Esta *función valor* presenta una discontinuidad en el vencimiento de la transacción, dependiendo de la cual se identifican vencimientos *duros* (*hard-real-time*), *blandos* (*soft-real-time*) o firmes (*firm-real-time*) (Figura 1).

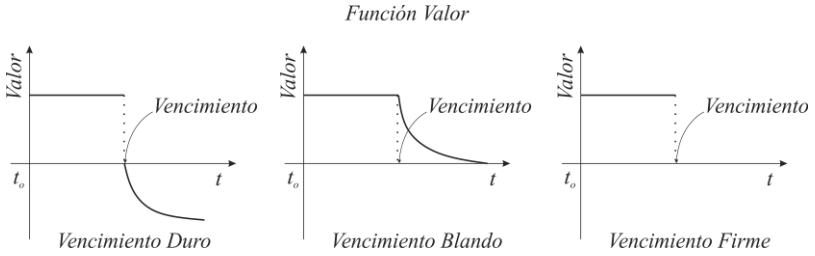


Figura 1. Función Valor.

Una decisión posible para el sistema, es que si al momento de fin de la *RTT* se obtiene una *función valor* negativa se debe deshacer la transacción (*rollback*).

Por otro lado, tampoco es posible finalizar satisfactoriamente una transacción cuyo *DataDeadline* expira antes que sea posible realizar el *commit*. Si se puede calcular el parámetro *estCompletionTime(t)* de una *RTT,* entonces es posible predecir si se podrá alcanzar el *commit* antes de que expire el intervalo de validez de los *RTD*. Si esto no es posible, el planificador puede analizar la posibilidad de retardar la transacción hasta que sus *RTD* se actualicen nuevamente. A este mecanismo se lo conoce como *Espera inducida* (*Forced Wait)*([19]). El planificador debe contemplar:

IF $dataDeadline_{RTT}(t) < estCompletionTime_{RTT}(t)$: WAIT UNTIL $dataDeadline_{RTT}(t)$

4 Framework RTT para Transacciones de Tiempo Real

La construcción de *RTDBS* requiere considerar los conceptos descriptos en el punto anterior. Resulta útil para los desarrolladores de aplicaciones contar con un marco de trabajo que facilite el diseño de *RTT* incluyendo atributos, servicios y reglas de validación de *consistencia temporal*. De esa manera, el desarrollador se puede enfocar más específicamente a la problemática propia del sistema a modelar.

En este trabajo se propone el *framework RTT*. Este es, un modelo orientado a objetos que contempla las características enunciadas para las *RTT* y que ofrece la información necesaria para que el *planificador de tiempo real* pueda implementar cualquiera de las políticas de planificación.

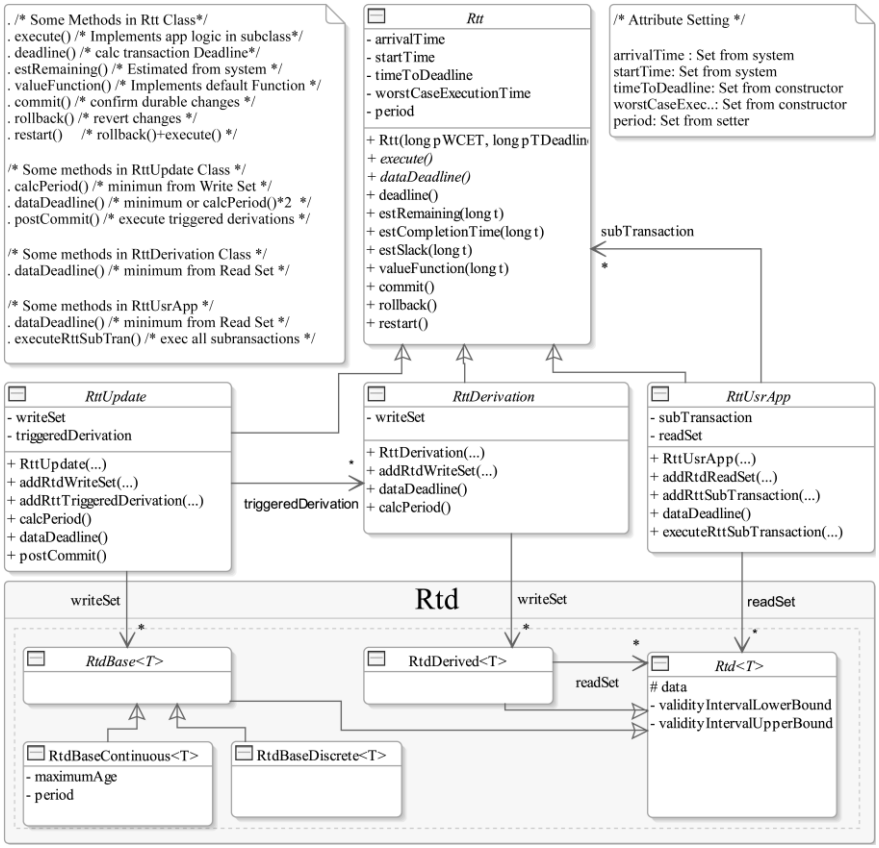


Figura 2. Framwork *RTT*

Para los datos de tiempo real se ha utilizado un paquete con el tipo de dato abstracto *RTD* ([13]), el cual caracteriza las diferentes clases de los *RTD*, encapsula la validación de *consistencia temporal* de los datos y calcula el *Datadeadline* de cada objeto. En la Figura 2 se describe el *framework RTT* en lenguaje *UML*.

El *framework* presenta cuatro clases abstractas. Una clase principal *Rtt* que define atributos y métodos comunes a todas las transacciones de tiempo real y tres subclases que implementan las particularidades de *RttUpdate*, *RttDerivation* y *RttUsrApp* como fue descrito en el punto 3.2. La subclase *RttUpdate,* permite asociar una lista de *RttDerivation* que serán disparadas (*triggered*) en el *postCommit*(). La subclase *RttUsrApp,* permite asociar una lista de subtransacciones *Rtt* que serán ejecutadas al invocar *executeRttSubTransaction*().

Los atributos, métodos y conjuntos (*Read Set* y *Write Set*) de cada clase fueron descriptos en 3.3 y se realizan aclaraciones adicionales en notas de la Figura 2.

El desarrollador de aplicaciones deberá implementar las *RTT* del sistema simplemente extendiendo la subclase que corresponda y definiendo la lógica de la transacción dentro del método *execute*(). Adicionalmente, se puede definir la *Función Valor* (*valueFunction*()) de acuerdo a si se desea implementar una *RTT* de vencimiento *duro*, *firme* o *blando*. Por defecto, *Rtt.valueFunction*() se define con vencimientos *firmes*. En la siguiente sección, se muestra el uso del *framework RTT* aplicándolo sobre un caso de ejemplo.

5 Aplicación en un caso de ejemplo

Se desea implementar un monitoreo de salud automatizado sobre una máquina *SCADA* (*Supervisory Control And Data Acquisition*). Se realizan diversas mediciones, entre ellas: la carga de trabajo (*measureVal*), la cual se calcula en función de la carga de CPU (*cpuWorkLoad*), la carga de memoria (*memWorkLoad*) y la tendencia de carga de los últimos minutos (*loadTrend*). Las mediciones tomadas se registran en la base de datos cada 20 segundos.

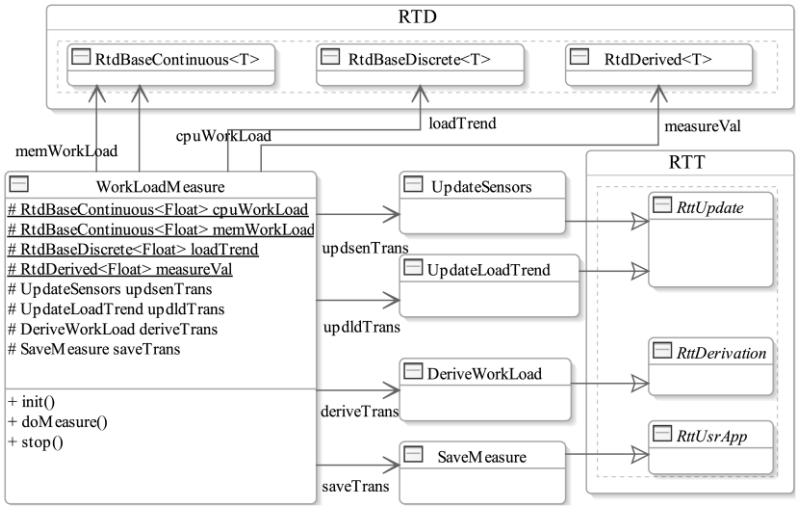


Figura 3. Aplicación del Framwork *RTT* para implementar sistema de mediciones

La carga de CPU y de memoria, se obtienen con sensados periódicos cada 4 y 6 segundos respectivamente. La tendencia de carga se obtiene desde la base de datos.

Para implementar este *RTDBS* se aplica el *framework RTT* y el diseño de transacciones se muestra en la Figura 3. En el diagrama se definen los *RTD* necesarios y cuatro *RTT* encargadas de obtener datos desde los sensores, calcular la medición y grabar en la base de datos. La clase *WorkLoadMeasure* es la encargada de automatizar la medición. En la Figura 4 se muestra la lógica aplicada:

```
init() :
 /*Define DTR para sensores */
 cpuWorkLoad = new RtdBaseContinuous<Float>(4 secs);
 memWorkLoad = new RtdBaseContinuous<Float>(6 secs);

 /*Define DTR derivado. Calcula la medición WorkLoad */
 measureVal = new RtdDerived<Float>();
 measureVal.addRtdReadSet(cpuWorkLoad);
 measureVal.addRtdReadSet(memWorkLoad);

/* Define transacción de Cálculo de derivaciones */
deriveTrans = new DeriveWorkLoad(pWCET→0.02 secs,
                                 pTDeadline→4 secs);
deriveTrans.addRtdWriteSet(measureVal);

/* Define transacción de Update de sensores */
updsenTrans = new UpdateSensors(pWCET →0.05 secs ,
                                pTDeadline→4 secs);
updsenTrans.addRtdWriteSet(cpuWorkLoad);
updsenTrans.addRtdWriteSet(memWorkLoad);
updsenTrans.addRttTriggeredDerivation(deriveTrans);

/* Tendencia de carga con dato de sensado discreto */
 loadTrend = new RtdBaseDiscrete<Float>();

 /* Define transacción de Update de sensado discreto */
 updldTrans = new UpdateLoadTrend(pWCET→0.1 secs,
                                  pTDeadline →4 secs);
 updldTrans.addRtdWriteSet(loadTrend);

 /* Define transacción de Grabar Medición */
 saveTrans = new SaveMeasure(pWCET →0.1 secs,
                             pTDeadline→20 secs);
 saveTrans.addRtdReadSet(measureVal);
 saveTrans.addRttSubTransaction(deriveTrans);

doMeasure() :
  /* Planifica Update de Sensores en RTT periódica*/
 updsenTrans.setPeriod(updsenTrans.calcPeriod());
 RTTScheduler.dispatch(updsenTrans, _PERIODIC);

 /* Planifica Registro de Medición en RTT periódica*/
 saveTrans.setPeriod(20 secs.);
 RTTScheduler.dispatch(saveTrans, _PERIODIC);
```

Figura 4. Métodos de la clase *WorkLoadMeasure*

El planificador de tiempo real *RTTScheduler* instancia las transacciones en el período indicado por el atributo *period* de la *RTT* pasada como parámetro y al llegar el momento de inicio, la ejecuta invocando al método *execute*().

6 Conclusiones y Trabajos Futuros

El trabajo presenta en detalle los conceptos de *consistencia temporal* en transacciones de tiempo real, que han sido presentados en diferentes trabajos del área. El objetivo principal fue elaborar una herramienta que guíe y simplifique el diseño de transacciones en un *Sistemas de Bases de Datos de Tiempo Real*. Consecuentemente, se desarrolló el *framework RTT*, que es un modelo orientado a objetos que incorpora clasificaciones, atributos y servicios necesarios, para implementar los tres tipos de *RTT* identificadas en el trabajo. Con esta herramienta, el desarrollador de aplicaciones se puede enfocar específicamente en el problema a resolver y dejar bajo la responsabilidad del *framework* y del *real-time scheduler,* lo referido a gestión de las transacciones de tiempo real. Finalmente y a modo de verificación, se aplica el *framework RTT* en la resolución de un problema concreto. No obstante, algunas cuestiones como control de concurrencia, tolerancia a fallos, ambientes distribuidos, infraestructuras de desarrollo, etc. no han sido consideradas. Estos temas se desarrollarán en futuros trabajos sobre esta temática.

Referencias

[1] R. Elmasri and S. Navathe, *Fundamentals of Database Systems 5/E*, 5/E ed.: Addison-Wesley, 2007.

[2] K. Ramamritham, "Real Time Databases," *International Journal of Distributed and Parallel Databases,* vol. 1, pp. 199-226, 1993.

[3] B. Purimetla*, et al.*, "Real-Time Databases: Issues and Applications," in. vol. ch.20, ed: in S.Son (ed.) Advances in Real-Time Systems, Prentice Hall, 1995.

[4] M. Xiong*, et al.*, "Maintaining Temporal Consistency: Issues and Algorithms," in *Proceedings of International Workshop on Real-Time Database Systems*, 1996, pp. 2-7.

[5] R. Abbott and H. Garcia-Molina, "Scheduling Real-Time Transactions: A Performance Evaluation," in *Proceedings of the 14th VLDB Conference*, 1988.

[6] J. A. Stankovic*, et al.*, "Misconceptions About Real-Time Databases," *IEEE Computer,* vol. 32, pp. 29-36, 1998.

[7] J. J. Prichard*, et al.*, "RTSORAC: A Real-Time Object-Oriented Database Model," *In The 5th International Conference on Database and Expert Systems Applications,* pp. 601-610, 1994.

[8] L. C. DiPippo and L. Ma, "A UML Package for Specifying Real-Time Objects," *Computer Standards & Interfaces* vol. 22, pp. 307-321, 2000.

[9] N. Idoudi*, et al.*, "Structural Model of Real-Time Databases: An Illustration," in *Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on*, 2008, pp. 58-65.

[10] K. Ben*, et al.*, "Maintaining temporal consistency of discrete objects in soft real-time database systems," *Computers, IEEE Transactions on,* vol. 52, pp. 373-389, 2003.

[11] S. Rekhis*, et al.*, "Modeling Real-Time applications with Reusable Design Patterns," *International Journal of Advanced Science and Technology,* vol. 22, pp. 71-86, 2010.

[12] A. Hala*, et al.*, "A General Framework for Modeling Replicated Real-Time Database," *International Journal of Electrical and Computer Engineering. Word Academy of Science, Engineering and Technology,* pp. 4-8, 2009.

[13] C. Buckle*, et al.*, "Abstract Data Type for Real-Time Database Systems," in *XVII Congreso Argentino de Ciencias de la Computación*, UNLP. La Plata, Argentina, 2011.

[14] Soparkar*, et al.*, *Time-Constrained Transaction Management: Real-Time Constraints in Database Transaction Systems*: Kluwer Academic Publishers, 1996.

[15] B. Adelberg*, et al.*, "Applying update streams in a soft real-time database system," *Proceedings of the 1995 ACM SIGMOD,* vol. 24, pp. 245-256, 1995.

[16] Y. Wei*, et al.*, "Maintaining Data Freshness in Distributed Real-Time Databases," presented at the Proceedings of the 16th Euromicro Conference on Real-Time Systems, 2004.

[17] C. L. Liu and J. W. Layland, "Scheduling Algorithms for Multiprogramming in a Hard Real-Time Environment," *Journal of the ACM,* vol. 20, pp. 46-61, 1973.

[18] E. D. Jensen*, et al.*, "A Time-Driven Scheduling Model for Real-Time Operating Systems," *Proceedings of Real-Time Systems Symposium,* pp. 112-122, 1985.

[19] M. Xiong*, et al.*, "Scheduling access to Temporal Data in Real-Time Databases," in *Real-Time Database Systems: Issues and Applications*, Bestavros*, et al.*, Eds., ed: Kluwer Academic Publishers, 1997, pp. 167-191.

Toolchain and workflow for the design of an ISO 11783-compatible ECU based on ISOAgLib

Joaquin Ezpeleta and Sebastián Rossi,

Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario,
Av. Pellegrini 250, S2000BTP Rosario, Argentina
ezpeleta@fceia.unr.edu.ar, srossi@inti.gob.ar

Abstract. This paper describes a basic toolchain for the design of an ISO 11783-compatible electronic control unit (ECU), from its conception to the implementation of a working embedded prototype, along with a suggested workflow for dividing application programming, mask design and hardware-related tasks in a debugging-friendly and time-efficient manner. The toolchain is centered on the open source ISOAgLib programming library distributed and maintained by OSB AG and the paper will refer to other specific tools and devices, but is otherwise intended to provide a general introductory overview of the process rather than focus on specific vendors or platforms.

Keywords: toolchain, workflow, ISO 11783, ISOBus, ISOAgLib, controller area network (CAN), embedded system.

1 Introduction

ISO 11783 [1 *et seq.*] –commonly known as ISOBus– defines a serial data network for agricultural or forestry equipment based on the CAN 2.0 B [4] protocol. It is intended to provide an interconnection system for on-board electronics. A typical ISOBus network is shown on Figure 1. A basic element of such network is the electronic control unit (ECU), which is defined in [1] as an 'electronic item consisting of a combination of basic parts, subassemblies and assemblies packaged together as a physically independent entity'.

This paper describes the process for creating one such ECU, from its design to the implementation of an embedded prototype. Specifically, it suggests a toolchain and a workflow for this process. The toolchain is based on the open-source library ISOAgLib and other related tools distributed and maintained by OSB AG [7], but alternatives are given wherever possible to provide maximum flexibility. Similarly, the process described is mostly platform-independent (both in terms of the operating system used for development on a desktop environment and in terms of the embedded hardware platform used for the actual implementation), but examples are given at various points for the purpose of clarity and reproducibility. For an implementation on specific hardware see, for example, [6].

Section 2 presents the suggested toolchain, along with explanations and brief examples; Section 3 describes the physical and virtual elements needed to test and use

software- and firmware-level applications; Section 4 presents the proposed workflow for an efficient and debugging-friendly task division using the tools and resources presented in Sections 2 and 3.

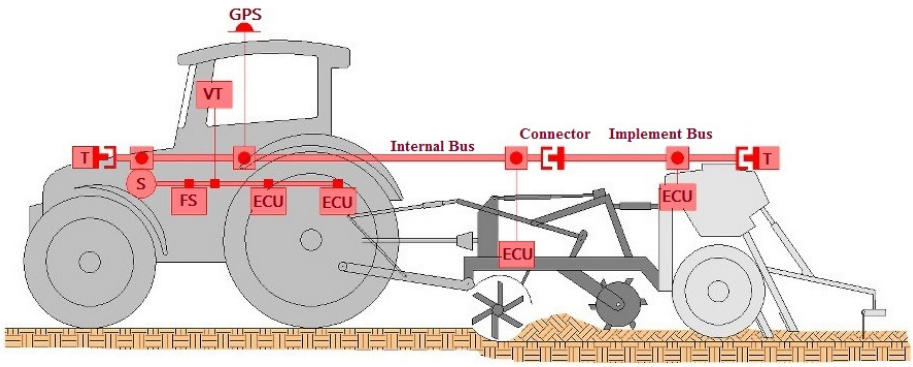


Fig. 1. Example ISO 11783 network on a tractor with two implements.

2 Toolchain

A summary of the suggested toolchain is shown graphically in Figure 2. It presents the tasks needed to design an ISOBus system with ISOAgLib, the tools needed for each task and the relationship between these tasks. As will be seen on Section 4, however, many of these tasks can often be undertaken independently, so a roughly left-to-right and top-to-bottom order will be followed.

2.1 Mask design and parsing

The first tool to be discussed is VT designer, which is a graphic interface for designing masks which are compatible with [3]. It is currently available from OSB AG in both a trial and a full version. It has a simple and intuitive interface and includes examples which can be modified to gain insight into its use. The purpose of this tool within the suggested toolchain is to take an abstract design concept for the necessary mask(s) and create files which can be further processed for use in the application. Given one or more manually-loaded masks (collectively, an object pool), VT Designer creates a .vtp project file and a number of .xml files which contain the actual objects and attributes.

An alternative for the use of VT Designer is PoolEdit, developed by Matti Öhman and Jouko Kalmari at Aalto University [9] and distributed under GNU General Public License (GPL). The former is discussed here for its greater affinity with the OSB AG toolchain, but the latter produces almost identical results.

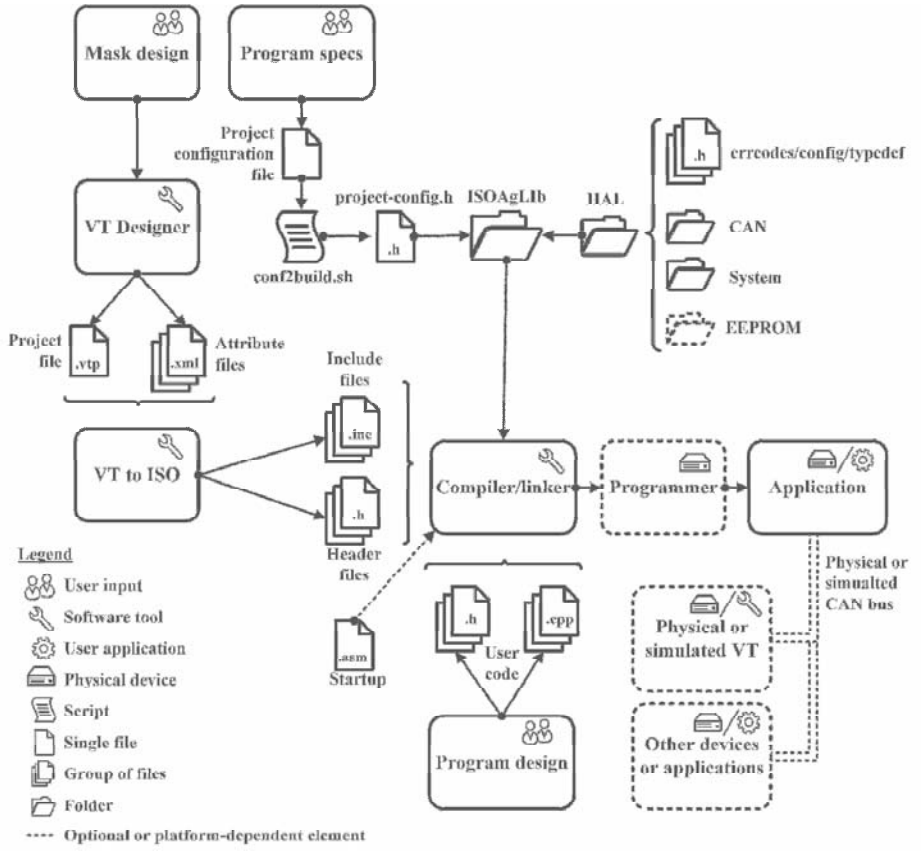


Fig. 2. Toolchain for the development of an ISO 11783-compatible application.

The .xml files output by VT Designer cannot typically be included directly into the source code of the application. A second tool is needed to transform the .xml files into source files which can be handled by the C++ compiler. An example of an application which serves this purpose is vt2iso.exe, available free of charge from OSB AG. It is a console application which essentially takes the .vtp and .xml files from the previous step and produces a group of .inc and .h files which can then be included directly by the compiler. Below is a basic use example on the Microsoft Windows command line, where myDisplay.vtp is a VT Designer project file. A number of optional arguments are also available but are not typically necessary.

```
C:\IsoAgLib\tools\vt2iso\bin>vt2iso.exe myDisplay.vtp
```

2.2 Project Configuration

Another step in the design process is defining project settings. These are to be entered manually in a configuration file and include basic parameters, such as the project name, the folder where the user and library files are stored, the number of CAN instances and the target platform (e.g. PC running Microsoft Windows, PC running Linux, embedded system, etc.), among others.

It is advisable to adapt the template (conf_template) included with ISOAgLib rather than enter the settings directly on a blank file, as the former approach will be both faster and less error-prone. In either case, the resulting file is to be processed by a shell script (conf2build.sh) to produce a configuration header (isoaglib_project_config.h) which can be included directly into the application.

The shell script is included with ISOAgLib and can be run natively in Linux and other UNIX systems. Microsoft Windows users will need to install MSYS/MINGW [11] or other similar software which makes it possible to run UNIX shell scripts under this OS. A typical command for executing the script is shown below, where myconfig is a copy of conf_template adapted to a specific project.

```
$ conf2build.sh myconfig
```

Also note that the resulting isoaglib_project_config.h header is included via the library header isoaglib_config.h and is not to be included directly (the preprocessor will issue an error if it is).

2.3 Hardware Abstraction Layer (HAL)

ISOAgLib includes a Hardware Abstraction Layer (HAL) which acts as an intermediary between the rest of ISOAgLib and the actual hardware on which it runs. This includes system startup, time tracking, power management, CAN communication, non-volatile storage and similar hardware-dependent functions, along with definitions for data types, error codes and configuration parameters.

The HALs for certain hardware platforms (most notably, PCs running Microsoft Windows or Linux) are already included with the library, but others must be programmed by the user (e.g. ARM-based MCUs). It is advisable to use an existing HAL as model when creating a new one. It should also be borne in mind that, when changing platforms, the project configuration described in 2.2 must be updated accordingly.

In embedded systems, it may also be necessary to include additional files outside the HAL, such as a startup file (see Figure 2, bottom left).

2.4 User Code

While ISOAgLib provides for most communication and compliance requirements, it is up to the user to design and program the actual application. The typical application includes initialization code which is run once at the start, a main application loop

which performs all communication and normal operation tasks and closing code which shuts down the application and all associated hardware in a controlled manner. In addition, on embedded systems, startup code is typically needed before the rest of the application can be executed.

The initialization code should perform the following basic functions: (a) initialize instances of the system, the scheduler, the bus, the monitor and the CAN controller; (b) declare or retrieve parameters needed for subsequent address claim; (c) register the necessary tasks on the scheduler, initialize the system components and register the object pool on the monitor; (d) transfer control to the main loop. In addition, it should perform any application-specific or hardware-dependent initialization functions (such as interrupt vector, watchdog, real-time clock, analog-digital converter or output configuration or pin remapping), although these can usually be called automatically during system initialization by including them in the HAL.

The main application loop should run after initialization and until the application needs to terminate for any reason. As far as the ISOAgLib library is concerned, the only requirement for the main application loop is that the time event of the scheduler instance be called within the time constraints defined in [1 *et seq.*], but other functions will be needed for any specific application.

Finally, the closing code unregisters all elements, deallocates memory, calls the close method for each of the active instances and takes all associated hardware to a safe condition (e.g. deenergized actuators).

Gaining insight into the use of ISOAgLib can be challenging given the lack of freely available training and tutorials. OSB AG offers customized workshops and training sessions which include examples, but the examples themselves cannot be bought alone. If workshops or training are not an option due to cost, examples are freely available for earlier versions of ISOAgLib (up to 2.2.1) from OSB AG's repository. These can be adapted to more current versions by following the changelog and inferring necessary changes in the user files, but this involves considerable guesswork and can be a time-consuming process.

2.4 Compilation, Linkage and Loading

The tools used for compilation, linkage and (in the case of embedded systems) loading are highly platform-dependent and are usually managed by a single IDE. General purpose compilers such as GCC [10] can be used for early stages of development where the application is to run entirely within the desktop environment, while vendor- or platform-specific compilers are generally needed to compile the application for specific hardware (e.g. armcc for ARM-based MCUs).

Remark 1. ISOAgLib is written in C++. This must be taken into account when selecting the product and compiler, as some products do not offer C++ compilers or do so at relatively high prices.

3 Hardware framework

Figure 3 shows an example interconnection of system elements. It is intended to provide a summary of the connection options available throughout the entire design process. In each stage of the development proces, however, not all of the elements are simultaneously necessary in general. For example, while initially programming and debugging the application, all work is typically done within a desktop environment without resorting to additional external hardware. Similarly, the process for preliminary design and implementation of sensors and actuators will not normally require access to a virtual terminal, whether physical or simulated.

The interconnection between system elements is done basically by means of three CAN buses, namely a physical bus, a socket bus and a proprietary virtual bus.

The physical bus is an actual wired bus complying with the requirements set forth in [2]. It includes a pair of data wires CAN_H and CAN_L and should be terminated with impedance-matching resistors.

Remark 2. While [2] defines a nominal characteristic bus impedance of 75 Ω, existing CAN hardware not developed specifically for ISOBus may use 120 Ω instead, as defined in [5] (the original Bosch document, [4], did not specify this and other electrical aspects of the physical layer).

Data transmission from and to the bus is normally done through CAN transceivers. These convert 3.3 V or 5 V logic values from the RX/TX pins on an MCU (or other levels from other hardware) to dominant and recessive bits on the CAN bus and vice versa. Examples of CAN transceivers are SN65HVD230 and MAX3051 for 3.3 V systems and SN65HVD255 and MAX3058 for 5 V systems.

The socket bus is an internal socket connection which emulates a CAN bus within a desktop environment. The ISOAgLib toolchain includes several similar tools for this purpose, one which is designed solely for interaction between several ISOAgLib applications within the desktop (can_server_simulating.exe) and others which additionally provide functionality for connection with vendor-specific hardware and software. For example, can_server_vector.exe and can_server_vector_xl.exe are designed for communicating with hardware and software from Vector Informatik GmbH [8], which is achieved by using Vector's CAN DLL library. The CAN server thereby acts as a bridge between the socket bus and the proprietary virtual bus, discussed below.

The proprietary virtual bus is an emulated CAN bus like the socket bus, and can be thought of as an extension of the latter. It is used for communication between proprietary hardware and software from a given vendor and the corresponding CAN server. In addition, by communicating with a CAN card, the proprietary virtual bus in turn enables access to the physical external bus. For example, Vector offers CANcardXL, a two-channel driver card which provides access to up to two external physical buses. Additionally, they offer an application by the name of CANoe which manages the card and also serves as a bus analyzer. When both CANcardXL and can_server_vector_xl.exe are running, the socket bus, the proprietary virtual bus and the physical bus merge, for all practical purposes, into a single CAN bus. This and other similar setups offer great versatility, as they make it possible for a number of

applications running either on the desktop or on an embedded system to communicate seamlessly with each other and with third-party devices, such as a virtual terminal. In addition, software like CANoe can emulate a virtual terminal, which is essential for running preliminary tests entirely within the desktop environment, with further tests on actual virtual terminals being done at a later stage of development.

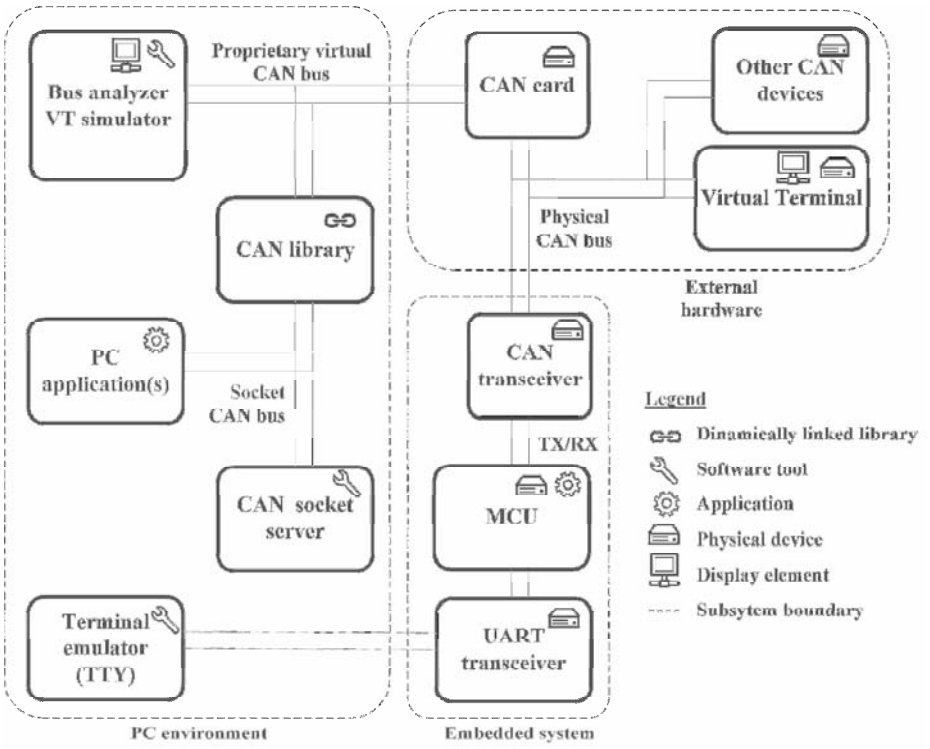


Fig. 3. Hardware framework.

In addition to the CAN buses, Figure 3 shows a UART connection between the MCU and a terminal emulator within the PC. This provides a simple means of transmitting debugging information between the PC and the MCU without loading the CAN bus, but is entirely optional.

4 Workflow

This Section describes a proposed workflow for the design process of an ISO 11783-compatible application. The basic stages in this process are application programming (with and without masks), mask design, hardware-related tasks and prototyping. A way of organizing these stages is shown in Figure 4. The basic concept is to run tasks in parallel in order to make better use of available time, hardware and

human resources. Time is better used as different teams can work on the different task simultaneously. In addition, each task can be assigned to a specialist in the corresponding field, such as a programmer for building the application and an electrical engineer for working wiring, sensors and actuators, making better use of available human resources. Finally, available hardware can be used more efficiently, as each task requires only some of the hardware tools. Some examples of this were discussed in the previous Section, which described a flexible hardware framework.

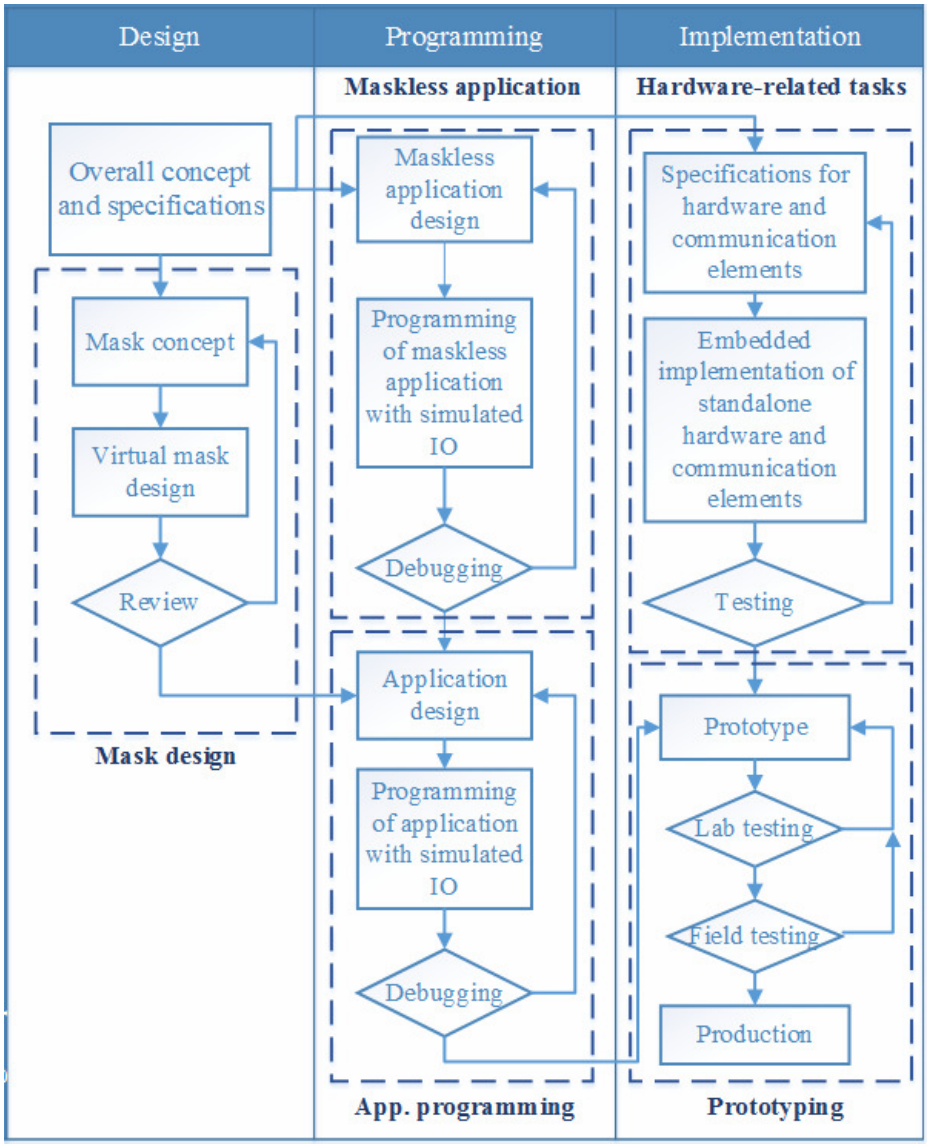


Fig. 4. Proposed workflow for the design process.

Another advantage of task division is to facilitate the debugging and troubleshooting process. When each task is undertaken individually, the possible error sources are confined to a subdomain of the entire system. For example, if the application is initially debugged using a virtual bus and simulated data (maskless application or application programming stage in Figure 4), no communication- or hardware-related errors arise and the debugging domain is thereby restricted only to the application itself.

The maskless application programming stage consists in the design and programming of a virtual application which does not resort to any virtual terminal (physical or otherwise) for display. Additionally, such application would be developed entirely within a desktop environment, without resorting to external data or elements and without being loaded onto an MCU or other embedded system. The application uses the PC HAL at this point. As the application will be isolated from data which it normally needs for its operation (data from sensors, other network devices or user input), simulated data can be used during this stage to examine the behavior of the application in different situations.

The mask design stage basically involves the task of designing one or more masks (an object pool) and parsing then into a format which can be handled directly by the compiler used for building the application. This stage can be completed mostly using the tools described in 2.1 above, although some previous design work is necessary to create a visual concept for the masks (colors, layout, shape and position of various elements, expected functionality, etc.).

The application programming stage is an extension of the maskless application programming stage with the addition of the masks designed in the mask design stage. The goal of this stage is to ensure that the application can upload the object pool to a simulated virtual terminal and interact with it (i.e. receive user input and make necessary changes on the elements of the object pool). Except for data flowing to and from the virtual terminal, the rest of the data used during this stage is still simulated. Similarly, the application continues to run within the desktop environment using a PC HAL.

The hardware-related tasks stage encompasses all hardware or platform-specific tasks. These basically include the design and implementation of actuators, drivers, sensors, data acquisition means, the creation of a HAL for ISOAgLib if one does not exist for the intended platform and the inclusion or programming of other platform-dependent elements which cannot be included within the HAL (such as a startup file). The creation of the HAL in turn involves implementing functions for CAN communication, configuration of interrupt vectors, watchdogs, real-time clock, analog-digital converters or outputs and pin remapping.

Finally, the application (included the masks) can be loaded onto the embedded hardware and combined with sensors, actuators, CAN peripherals and other hardware elements to produce a working embedded prototype. This prototype can then be tested for minor bugs or problems within a laboratory environment (possibly using a simulated VT) to create a final version of the device, which can be further tested with actual equipment on field prior to its commercial production.

References

1. ISO 11783-1:2007, Tractors and machinery for agriculture and forestry — Serial control and communications data network — Part 1: General Standard for mobile data communication, International Organization for Standardization, Geneva (2007)
2. ISO 11783-2:2002, Tractors and machinery for agriculture and forestry — Serial control and communications data network — Part 2: Physical Layer, International Organization for Standardization, Geneva (2002)
3. ISO 11783-6:2004, Tractors and machinery for agriculture and forestry — Serial control and communications data network — Part 6: Virtual Terminal, International Organization for Standardization, Geneva (2004)
4. CAN Specification Version 2.0 Part B, Robert Bosch GmbH, Stuttgart (1991)
5. ISO 11898-2:2003, Road vehicles — Controller area network (CAN) — Part 2: High-speed medium access unit, International Organization for Standardization, Geneva (2003)
6. Tumenjargal, E., Badarch, L., Kwon, H., Ham, W.: Embedded software implementation system for a human machine interface based on ISOAgLib. Journal of Zheijang University (2013)
7. OSB AG, http://www.osb-ag.com/osb-ag.html
8. Vector Informatik GmbH, http://vector.com
9. PoolEdit - Open Source XML ISO 11783 User Interface Editor, http://autsys.aalto.fi/en/Farmix/PoolEdit
10. GCC, the GNU Compiler Collection, http://gcc.gnu.org
11. MSYS, http://www.mingw.org/wiki/MSYS

Controlador neuronal incremental aplicado a un mezclador de flujos

Sergio L. Martínez[1], Enrique E. Tarifa[1,2] & Samuel Franco Dominguez[1]

(1) Facultad de Ingeniería, Universidad Nacional de Jujuy,
Gorriti 237, S. S. de Jujuy, Jujuy, Argentina
{smartinez, eetarifa, sfdominguez}@fi.unju.edu.ar
(2) Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).
eetarifa@arnet.com.ar

Resumen. En este trabajo se diseña e implementa un controlador tipo MIMO basado en redes neuronales artificiales, aplicado a un modelo de mezclador de corrientes líquidas, configurado sobre el entorno de simulación gráfica de Matlab®. Se destaca una particular metodología de entrenamiento del controlador neuronal, basada en un punto de operación genérico asociado a un reducido entorno del espacio de datos definidos en forma incremental, permitiendo una importante reducción de datos de aprendizaje y esfuerzo computacional. El desempeño del controlador es comparado con otro controlador neuronal similar configurado bajo un proceso de entrenamiento estándar. Los resultados obtenidos permiten apreciar las ventajas del método de control incremental.

Palabras clave. Redes neuronales. Mezclador de flujos. Control. Simulación.

1 Introducción

Los dispositivos mezcladores de flujos son sistemas adicionales indispensables en muchos procesos industriales y son objeto de estudio y aplicaciones en diversas investigaciones [1], [2], [3]. Usualmente se complementan con sistemas de control que requieren correcciones permanentes, que según la complejidad de la dinámica del proceso podrían ser realizadas por un operario experto; o por un sistema de control automático.

En este trabajo se diseña e implementa un controlador tipo MIMO (*Multiple-input Multiple-output*) basado en redes neuronales, aplicado al modelo de un mezclador de corrientes líquidas, configurado sobre el entorno Simulink® de Matlab®. Se destaca una particular metodología de entrenamiento del controlador, basada en un punto de operación genérico que incursiona sobre un reducido entorno definido en forma incremental, permitiendo una importante minimización de datos requeridos para el aprendizaje de la red neuronal. El desempeño del controlador es comparado con otro controlador neuronal similar configurado bajo un proceso de entrenamiento estándar. Los resultados obtenidos permiten apreciar las ventajas del método de control incremental.

2 Proceso a controlar

Los mezcladores de flujos o de caudales (*flow mixers*), utilizados en muchos procesos industriales, suelen estar sometido a dinámicas exigentes, requiriendo la asistencia de sistemas de control automáticos. Un sistema de este tipo, cuyo esquema y modelo gráfico se muestran en la Fig. 1, es utilizado en este trabajo como sistema a controlar.

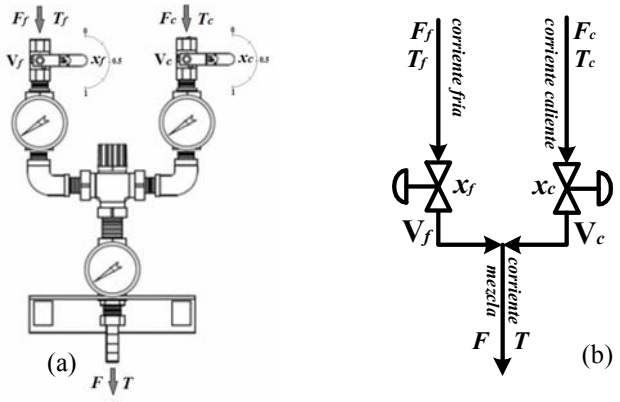


Fig. 1. Mezclador de caudales en línea. (a) esquema físico, (b) modelo.

Las entradas reciben las corrientes para la mezcla con caudales y temperaturas preestablecidos, y en la salida se obtiene una nueva corriente con propiedades específicas (caudal, presión, temperatura, composición).

2.1 Modelo del mezclador en línea

El modelo para el mezclador de flujos en línea se puede representar con el siguiente conjunto de ecuaciones:

$$F = x_{\mathrm{f}} F_{\mathrm{f}} + x_{\mathrm{c}} F_{\mathrm{c}} . \tag{1}$$

$$T = \frac{x_{\mathrm{f}} F_{\mathrm{f}} T_{\mathrm{f}} + x_{\mathrm{c}} F_{\mathrm{c}} T_{\mathrm{c}}}{x_{\mathrm{f}} F_{\mathrm{f}} + x_{\mathrm{c}} F_{\mathrm{c}}} . \tag{2}$$

$$0 \le x_{\mathrm{f}} \le 1 \text{ y } 0 \le x_{\mathrm{c}} \le 1 \tag{3}$$

donde la corriente de entrada (fría), tiene un caudal máximo F_{f} y una temperatura T_{f}. Esta corriente es regulada por la apertura x_{f} de la válvula $\mathrm{V_f}$. La otra corriente de entrada (caliente), tiene un caudal máximo F_{c} y una temperatura T_{c}. Esta corriente es regulada por la apertura x_{c} de la válvula $\mathrm{V_c}$. Las aperturas x_{f} y x_{c} incursionan en el intervalo (0, 1), donde 0 corresponde a la válvula completamente cerrada y 1 a la válvula completamente abierta, de modo que permite pasar la totalidad del caudal asignado. La corriente mezcla presenta a la salida un caudal F y una temperatura T.

2.2 Modelo inverso del mezclador en línea

Invirtiendo el modelo planteado, se obtiene un sistema de control ideal MIMO que proporciona las aperturas de las válvulas (x_f y x_c), para un caudal de referencia (*set-point*) F_{sp} y para una temperatura de referencia T_{sp} preestablecidos. El modelo matemático asociado a este sistema inverso está formado por las ecuaciones siguientes:

$$x_f = \frac{F_{sp}\left(T_c - T_{sp}\right)}{F_f\left(T_c - T_f\right)}. \tag{4}$$

$$x_c = \frac{F_{sp}\left(T_{sp} - T_f\right)}{F_c\left(T_c - T_f\right)}. \tag{5}$$

$$T_f \le T_{sp} \le T_c \text{ y } 0 \le F_{sp} \le F_{max} \tag{6}$$

donde F_{max} es el máximo valor que puede adoptar F_{sp} para un dado T_{sp}. El caudal F_{max} se obtiene a través de un problema de optimización, cuyo resultado se muestra en (7):

$$F_{max} = \begin{cases} \dfrac{F_f\left(T_c - T_f\right)}{T_c - T_{sp}}. & \text{si } \; T_f \le T_{sp} \le \dfrac{F_f T_f + F_c T_c}{F_f + F_c}. \\ \dfrac{F_c\left(T_c - T_f\right)}{T_{sp} - T_f}. & \text{en otro caso} \end{cases} \tag{7}$$

3 Sistemas de control

3.1 RNA como procesadores no lineales

Las redes neuronales artificiales (RNA), son modelos matemáticos que emulan –a nivel básico– la actividad del cerebro humano, dotados de la capacidad de aprender, "memorizar" y generalizar la información aprendida, con elevada tolerancia al ruido. Desde un punto de vista general, las RNA se especializan en descubrir y asociar patrones de entrada–salida con relación lineal o no lineal, según sea su arquitectura, configuración de las neuronas y proceso de aprendizaje [4], [5]. La Fig. 2 presenta una arquitectura típica de red neuronal feedforward de tres capas, con M neuronas de entrada, L neuronas en la capa oculta y N neuronas de salida.

La primera capa de la RNA –de entrada– se configura con unidades ficticias que distribuyen las señales hacia la capa oculta. La salida de cada neurona de esta capa responde a la siguiente ecuación:

$$y_l = g\left(-w_{0l} + \sum_{m=1}^{M} w_{lm}\, x_m\right) \quad \text{con } l = 1,\ldots,L. \tag{8}$$

donde x_m es la m-ésima componente del patrón de entrada; w_{lm} es el peso de conexión entre la neurona l de la capa oculta y la neurona m de la capa de entrada; w_{0l} es el peso de ajuste (bias) de las neuronas de la capa oculta; $g(\cdot)$ es la función de transferencia de las neuronas ocultas y y_l es la salida de la l-esima neurona oculta. La salida de cada neurona de la última capa se modela con la ecuación (9):

$$z_n = h\left(-v_{0n} + \sum_{l=1}^{L} v_{nl}\ y_l \right) \quad \text{con } n = 1,...,N\,. \tag{9}$$

donde y_l es el valor de salida de las neuronas de la capa anterior; v_{nl} es el peso de conexión entre la neurona n de la capa de salida y la neurona l de la capa oculta; v_{0n} es el peso de ajuste (*bias*) de las neuronas de la capa de salida; $h(\cdot)$ es la función de transferencia de las neuronas de salida y z_n es la n-ésima componente del patrón de salida Z.

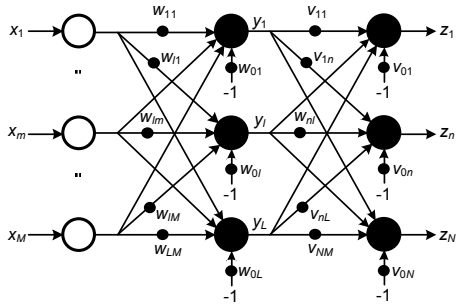


Fig. 2. Red feedforward tricapa genérica.

3.2 Controlador neuronal incremental

Se pueden adoptar dos enfoques generales para la implementación de controladores neuronales. En un primer enfoque, la RNA tiene como entrada tanto las variables controladas como los correspondientes valores de referencia, mientras que las salidas son las variables manipuladas (F, T, F_{sp} y T_{sp} en este caso); mientras que las salidas son las variables de control (x_f y x_c). Si se tiene algún conocimiento previo del sistema –tal como el modelo del proceso, secuencias históricas o muestras de ejemplo con sus correspondientes valores de salida–, se puede utilizar para el entrenamiento de la red. En este caso, la RNA aprenderá a asociar cada patrón de entrada con la salida correspondiente, sobre la totalidad del espacio de datos del sistema, dentro de un nivel de error predeterminado.

Aunque este es un enfoque clásico y ampliamente utilizado, tiene algunos inconvenientes que pueden considerarse críticos. El primer inconveniente surge por la gran cantidad de información que se requiere para cubrir adecuadamente el espacio de datos del sistema, que en muchas ocasiones no están disponibles. Otro inconveniente destacable es la compleja arquitectura que puede alcanzar la RNA, eventualmente con una importante cantidad de unidades neuronales internas, demandando un elevado esfuerzo computacional, que en ciertos casos obliga a aplicar técnicas adicionales para reducir el flujo de información a costa de un incremento en el error de aprendizaje [6].

En un segundo enfoque para implementar una RNA como controlador, se puede considerar la operación de la red bajo un esquema incremental. Para ello, se configura la RNA para que las entradas sean los errores de las variables controladas (e_F y e_T), y las salidas sean las correcciones que deben realizarse sobre las variables manipuladas (Δx_f y Δx_c). Estas nuevas variables se definen como:

$$e_F = F_{sp0} - F \qquad e_T = T_{sp0} - T\,. \tag{10}$$

$$\Delta x_f\left(t\right) = x_f\left(t\right) - x_f\left(t - \Delta t\right) \tag{11}$$

$$\Delta x_c\left(t\right) = x_c\left(t\right) - x_c\left(t - \Delta t\right) \tag{12}$$

donde t es el tiempo de simulación, y Δt es el paso de simulación. Con este cambio de variables, el punto de operación deseado es aquel que anula los errores. Cualquier desviación de este punto de operación, requiere que el controlador realice correcciones sobre las variables controladas. Es decir, cuando ambas entradas de la RNA sean nulas, sus salidas también lo serán y no es necesaria ninguna acción de control. En cambio cuando una o ambas entradas dejen de ser nulas, las salidas de la RNA deberán proveer las correcciones necesarias sobre las variables de control.

La conducta descripta es independiente del punto base que se tome. Más aún, si la superficie de control es lineal, las magnitudes de las correcciones serán independientes de la posición del punto base.

En este nuevo esquema de control, no es necesario entrenar a la RNA en toda la región de control. Concretamente, bajo este enfoque se establece un punto de operación genérico que representa la situación de referencia general de las variables de control y se definen las condiciones de operación sobre un entorno, mediante valores incrementales a partir del punto de operación establecido. En la Fig. 3 se muestra un espacio de datos bidimensional genérico con sus límites, un punto de operación base definido y el entorno asociado con los respectivos incrementos de las variables de control.

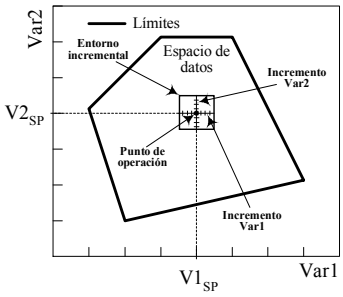


Fig. 3. Punto de operación y entorno incremental para un proceso genérico.

La adecuada implementación de la estrategia de control propuesta requiere algunas consideraciones adicionales. Por un lado está la selección del punto base, cuya ubicación no debe sobrepasar los límites del espacio de datos, y debería ser el punto

establecido por las condiciones nominales de diseño del sistema. Por otro lado, está la determinación de la extensión del entorno alrededor del punto base, que debe ser suficientemente extensa como para capturar la naturaleza de la superficie de control.

Un tercer criterio a considerar es el intervalo de subdivisión del entorno de operación: incrementos muy grandes no permiten capturar las variaciones del entorno de la superficie de control, generando errores considerables en otras ubicaciones del punto base; incrementos muy pequeños producirían gran cantidad de datos innecesarios por estar más allá de la sensibilidad de los componentes del sistema.

4 Desarrollo experimental

El modelo experimental del mezclador de flujos (Fig. 1) se ha implementado inicialmente para operación manual sobre el entorno Simulink® de Matlab® (Fig. 4) y se ha instanciado con los siguientes parámetros:

- Corriente fría: caudal de entrada $F_f = 100$ l/min, temperatura $T_f = 25$ ºC.
- Corriente caliente: caudal de entrada $F_c = 100$ l/min, temperatura $T_c = 70$ ºC.

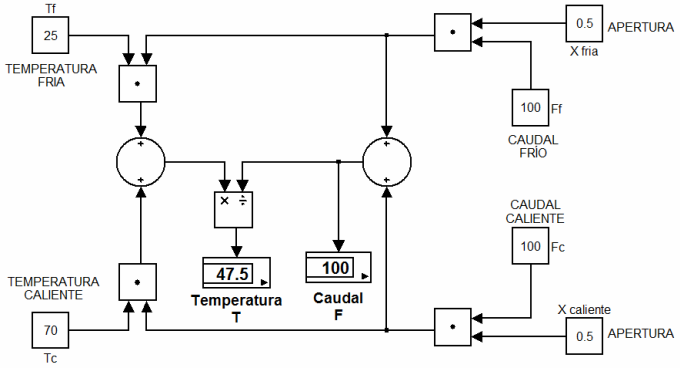


Fig. 4. Modelo Simulink® del mezclador de flujos en línea.

4.1 Generación de datos

Para determinar el punto base de operación se utilizó el modelo directo del mezclador, definido por las ecuaciones (1) a (3), donde las variables de entrada son las aperturas x_f y x_c, y las variables de salida son el caudal F y temperatura T de la mezcla. Evaluando este modelo para el caso en que ambas válvulas están abiertas a la mitad ($x_f = x_c = 0.5$), se obtiene a la salida un caudal $F = 100$ l/min y una temperatura $T = 47.5$ ºC. Este punto fue tomado como base para el proceso de entrenamiento de la RNA.

A continuación se estableció el entorno de operación en ±10% para ambas variables y la variación incremental ΔF y ΔT en el 1%, considerando que es ésta la mínima resolución del sistema. De esta manera, cada punto del entorno de operación queda definido de la siguiente forma:

$$F_{\mathrm{sp}i} = F_{\mathrm{sp}0} + (i-11)\Delta F\, F_{\mathrm{sp}0} \quad \text{con} \quad i = 1,\ldots,21$$
$$F_{\mathrm{sp}}^{\min} = 90 \ \text{l/min} \qquad F_{\mathrm{sp}}^{\max} = 110 \ \text{l/min} \tag{13}$$

$$T_{\mathrm{sp}j} = T_{\mathrm{sp}0} + (j-11)\Delta T\, T_{\mathrm{sp}0} \quad \text{con} \quad j = 1,\ldots,21$$
$$T_{\mathrm{sp}}^{\min} = 42.75\,^{\circ}\mathrm{C} \qquad T_{\mathrm{sp}}^{\max} = 52.25\,^{\circ}\mathrm{C} \tag{14}$$

El entorno de operación queda particionado en 21 intervalos, obteniéndose un total de 441 puntos de operación incrementales alrededor del punto base, como se esquematiza en la Fig. 5.

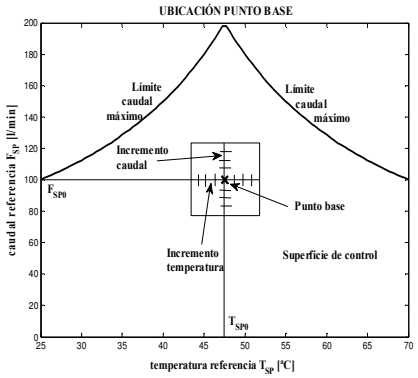


Fig. 5. Posición de punto base y entorno incremental del mezclador de flujos.

Definido el entorno de operación a estudiar, se utilizó el modelo inverso, ecs. (4) a (6), para obtener las acciones de control requeridas para cada uno de los puntos incluidos en el entorno del punto base. En este modelo inverso, las variables de entrada son el caudal de referencia (F_{sp}) y la temperatura de referencia (T_{sp}); mientras que las variables de salida son las aperturas x_{f} y x_{c} que permitirán alcanzar el estado deseado. Conocidas las acciones de control necesarias para llegar a cada punto del entorno de operación, se las expresa como cambios requeridos en función de los errores medidos:

$$e_{\mathrm{F}i} = F_{\mathrm{sp}0} - F_{\mathrm{sp}i} = -(i-11)\Delta F\, F_{\mathrm{sp}0} \tag{15}$$

$$e_{\mathrm{T}j} = T_{\mathrm{sp}0} - T_{\mathrm{sp}j} = -(j-11)\Delta T\, T_{\mathrm{sp}0}\,. \tag{16}$$

$$\Delta x_{\mathrm{f}ij} = x_{\mathrm{f}ij} - 0.5. \qquad \Delta x_{\mathrm{c}ij} = x_{\mathrm{c}ij} - 0.5. \tag{17}$$

4.2 Diseño del controlador neuronal incremental

A partir de una arquitectura neuronal feedforward, los parámetros más significativos a considerar son el número de capas y las cantidades de neuronas por capa. La dimen-

sionalidad de las capas de entrada y salida queda definida por las características del problema, siendo cada una bidimensional, en este caso. De acuerdo con la interpretación del teorema de aproximación de Cybenko [7], una capa oculta con una cantidad finita de unidades con funciones monótonas crecientes, es suficiente para cualquier mapeo no lineal de entrada-salida con un nivel de error suficientemente bajo.

Luego, el parámetro de mayor criticidad es la cantidad de neuronas ocultas encargadas del procesamiento interno; una cantidad insuficiente de neuronas ocultas puede impedir alcanzar el nivel de error deseado, mientras que una cantidad excesiva puede disminuir la capacidad de generalización de la red. En muchos casos, la determinación de este parámetro se realiza en forma experimental, aunque existen algunas heurísticas dedicadas a la determinación de la cantidad de neuronas ocultas, que aunque no son matemáticamente rigurosas, pueden producir buenas aproximaciones.

Así por ejemplo, se puede citar a la regla de la pirámide geométrica, donde el número de neuronas ocultas ($N^{(o)}$) se obtiene como una progresión geométrica entre el número de neuronas de entrada ($N^{(e)}$) y el número de neuronas de salida ($N^{(s)}$), de la forma $N^{(o)} = \sqrt{N^{(e)}.N^{(s)}}$ [8]. La regla de las capas entrada-oculta establece que el número de neuronas ocultas no debe superar dos o tres veces la cantidad de neuronas de entrada [9]. Otra regla práctica, utilizada por Goethals *et al*. [10], sugiere que el número de neuronas ocultas ($N^{(o)}$) se relaciona con el número de neuronas de entrada ($N^{(e)}$) de la forma $N^{(o)} = 2 \times N^{(e)} + 1$.

Bajo las consideraciones anteriores, se definió una RNA feedforward con arquitectura 2+5+2 para actuar como un controlador neuronal incremental tipo MIMO (2 entradas y 2 salidas), con las siguientes características:

- 2 neuronas en la capa de entrada.
- 5 neuronas en la capa oculta. Función de transferencia tangente sigmoide.
- 2 neuronas en la capa de salida. Función de transferencia lineal.

Con los parámetros anteriores definidos, la RNA fue entrenada con el algoritmo *backpropagation* obteniéndose los siguientes resultados:

- Cantidad de iteraciones: 500.
- Entrenamiento con algoritmo *backpropagation*, variante LM.
- Error cuadrático medio (ECM) de entrenamiento: 2.11×10^{-10}.

4.3 Configuración y operación del sistema

El modelo experimental del sistema completo (controlador–planta), fue configurado en el entorno de simulación gráfica de Matlab® (Fig. 6). En este modelo, el sistema a controlar se incorpora como un bloque que recibe los parámetros de caudales de entrada (F_f y F_c), de temperatura (T_f y T_c) de tales caudales y las variables de control (x_f y x_c), produciendo las variables de salida caudal (F) y temperatura (T).

El controlador neuronal, recibe a la entrada el error de caudal ($errF = F_{sp} - F$) y el error de temperatura ($errT = T_{sp} - T$), ambos modulados por limitadores que mantienen a los valores incrementales dentro del entorno definido, mejorando la estabilidad del sistema; y genera las variaciones de apertura de la válvula fría (Δx_f) y de la válvula caliente (Δx_c). Estas variaciones se componen con las aperturas del estado anterior ($x_f(k) = x_f(k-1) + \Delta x_f(k) \parallel x_c(k) = x_c(k-1) + \Delta x_c(k)$) para ser realimentadas al mezclador de flujos, cerrando el lazo de control.

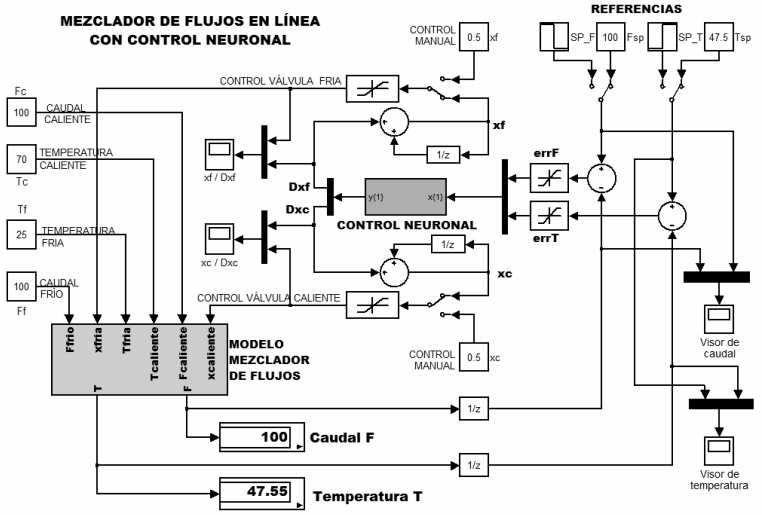


Fig. 6. Modelo experimental del mezclador de flujos con control neuronal.

4.4 Prueba del sistema

Para comprobar la operación del sistema, se aplicaron diferentes condiciones en los parámetros de referencia (*setpoints*). En el primer caso, los parámetros de referencias requirieron variaciones abruptas para el caudal y la temperatura como muestra la Fig. 7.

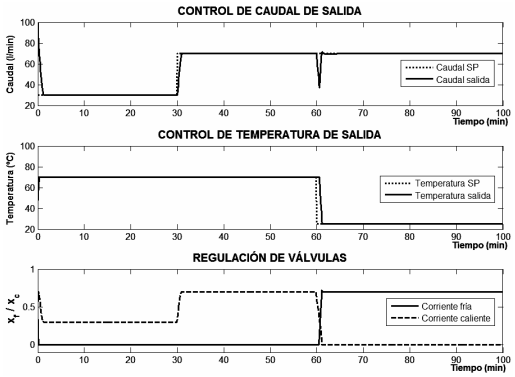


Fig. 7. Control de caudal y temperatura de salida para referencia tipo escalón.

Se observa que el sistema ejecutó una muy buena acción de control para responder a las variaciones abruptas de los valores de referencia de caudal y temperatura.

En el segundo caso, el parámetro de referencia caudal (F_{sp}) exigió una variación sinusoidal suave mientras que la temperatura (T_{sp}) debió mantenerse en un valor constante (Fig. 8). En este caso que la variación exigida por la referencia del caudal (F_{sp}) es correctamente seguida por la planta, mientras la temperatura se mantiene constante en $T = 47.5$ ºC como requiere la referencia T_{sp}.

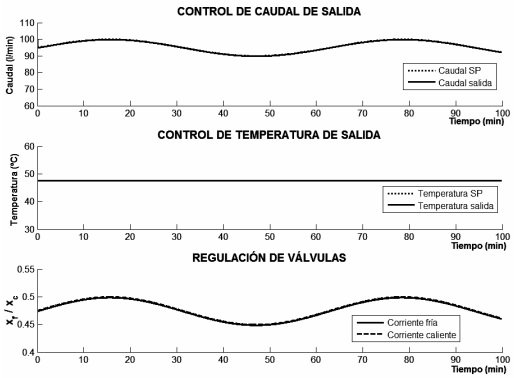


Fig. 8. Control de caudal y temperatura de salida para referencia tipo sinusoidal.

En el tercer caso, se provocó una perturbación sinusoidal del 2% en el parámetro caudal de corriente caliente (F_c), al mismo tiempo que se aplicó una variación lineal tipo rampa a la temperatura de referencia (T_{sp}), que se inició en 25 ºC, obteniéndose los resultados mostrados en la Fig. 9.

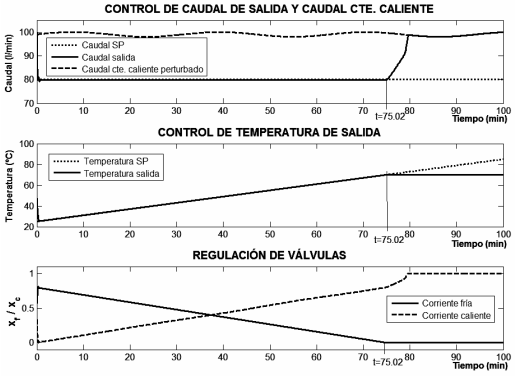


Fig. 9. Control de caudal y temperatura de salida para perturbación en caudal.

En este caso la perturbación fue adecuadamente absorbida por el controlador, hasta el instante $t = 75.02$ min, donde el controlador se satura (la válvula de la corriente fría se cierra totalmente, $x_f = 0$) y la temperatura de salida se estabiliza en $T = 70$ ºC.

El desempeño del controlador neuronal incremental (CN incremental) desarrollado se ha comparado con un controlador neuronal equivalente (CN estándar), con entrenamiento clásico, para la misma planta y con idénticos parámetros [6]. La comparación de los parámetros más representativos se observan en la Tabla 1.

Los datos de esta tabla, evidencian una indiscutible ventaja del controlador neuronal incremental en varios aspectos, tales como una arquitectura más simple, menor cantidad de variables de entrada, gran reducción en la cantidad de datos –sin necesidad de preprocesamiento previo–, menor tiempo de entrenamiento y mejora en el error general de aprendizaje.

Tabla 1. Parámetros de comparación entre controladores equivalentes.

Parámetro	CN incremental	CN estándar
Variables de entrada	2	4
Variables de salida	2	2
Cantidad de capas	3	3
Neuronas ocultas	5	10
Patrones de entrenamiento	2×441	4×5034
ECM de entrenamiento	2.11×10^{-10}	4.83×10^{-9}
Tiempo de entrenamiento	36 seg	1 min 47 seg

5 Conclusiones

Se ha diseñado e implementado en un sistema de simulación gráfica, un controlador neuronal MIMO configurado para trabajar por incrementos a partir de un entorno genérico predefinido, aplicado a un mezclador en línea de corrientes líquidas.

La capacidad del conjunto controlador–planta se ha comprobado experimentalmente bajo diversas condiciones, tales como variaciones abruptas y oscilantes de los parámetros de referencia y perturbaciones de los parámetros de la planta, mostrando un muy buen desempeño del sistema de control neuronal propuesto.

El controlador neuronal incremental se ha comparado con un controlador neuronal estándar equivalente aplicado a la misma planta, poniéndose en evidencia las ventajas en diseño, entrenamiento y operación del modelo incremental.

6 Referencias

[1] Palencia Díaz A. Estudio de Diferentes Estrategias de Control para un Tanque de Mezclado: PID, Control de Matriz Dinámica y Lógica Difusa. Prospect, 8(1), pp. 43-51, (2010)

[2] Mohamed Sultan M., Sha A.S., David C.O.: Controllers optimization for a fluid mixing system using metamodeling approach. In: International Journal of Simulation Modelling, 8(1), pp. 48–59, (2009)

[3] Yan Deng S.: Nonlinear & Linear MIMO Control of an Industrial Mixing Process. Master Thesis, McGill University, Montreal, Canada, (2002)

[4] Galushkin A.I.: Neural Networks Theory. Springer, New York (2007)

[5] He X., Xu S.: Process Neural Networks - Theory and Applications. Springer-Verlag, Berlín (2009)

[6] Martínez S.L., Tarifa E.E., Sánchez Rivero V.D.: Control neuronal tipo MIMO aplicado a un mezclador de corrientes líquidas. En: Investigaciones en Facultades de Ingeniería del NOA, T2, pp. 865-874. Catamarca, Argentina (2011)

[7] Poznyak A.S., Sanchez E.N., Yu W.: Differential Neural Networks for Robust Nonlinear Control. World Scientific Publishing, Singapore (2001)

[8] Blum A.: Neural Networks in C++: An Object-Oriented Framework for Building Connectionist Systems. John Wiley & Sons Inc, New York (1992)

[9] Berry M. J. A., Linoff G. S.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. John Wiley & Sons, New York (2011)

[10] Goethals P.L., Dedecker A.P., Gabriels W., Lek S., De Pauw N.: Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Springer - Aquatic Ecology, V. 41, pp. 491-508 (2007)

A Fault Resilience Tool for Embedded Real-Time Systems

Franklin Lima Santos[1], Flávia Maristela Santos Nascimento[2]

[1] Federal University of Bahia
Department of Electric Engineering
franklin_lima@ieee.org
[2] Federal Institute of Bahia
Department of Electro-Electronic Technologies
flaviamsn@ifba.edu.br

Abstract. Real-time systems have been used in many different areas such as medicine, multimedia and mechatronics. For such systems, it is important to meet both logical and timing requirements, since a malfunction may have undesired consequences. In this paper, we developed a simulation tool in MATLAB® environment to deal with fault-tolerant real-time scheduling under Rate Monotonic scheduling policy, so that errors consequences can be envisioned, before system is put on operation.

Keywords: Real-time systems, fault-tolerance, simulation, MATLAB®.

1 Introduction

Over the past decades computer designers focused their attention on developing what they considered a perfect computer project: computers had to be small, fast and cheap [12]. Indeed, their effort in reaching more performance at low cost and minimum size contributed remarkably for recent technological advances, especially those related to hardware improvements. The remarkable growth of electronic devices and computing systems in our daily activities has been boosting mechatronics, as a subarea of automation due to its ability of integrating electronic components and systems [4, 7]. The main elements of a mechatronic system can be observed in Figure 1.

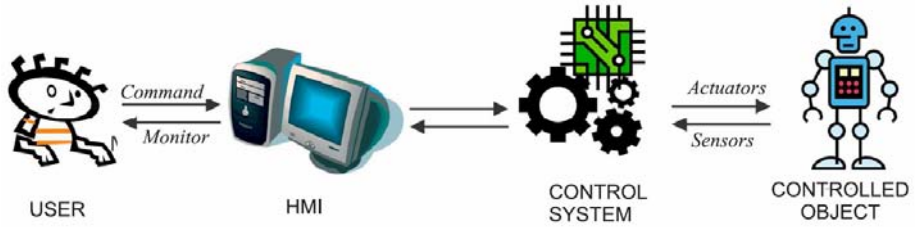


Fig. 1. Components of a mechatronic system [10].

The *user* is the entity responsible for monitoring, super visioning and controlling the *controlled object*, which may be an airplane board control or an industrial plant, for example. *Controlled objects* are usually manipulated through *human-machine interface* (HMI), which is the element responsible for (i) translating control information to the user and (ii) allowing an interface between users and controlled objects. The *control system* is an interactive computer system that enables monitoring and changing the state of a controlled object, which is done through sensors and actuators [8].

The evolution of computer systems also allowed systems designers to focus on modeling, designing and implementation aspects of such systems aiming at developing applications with differentiated performance skills. At the same time, miniaturization of electronic components allowed computers to evolve from simply terminals to host control systems. For some of these applications correctness were not only associated with logical, but also with timing requirements. Indeed, systems in which correctness is associated not only with producing logically correct results, but also with the time at which such results are produced (timeliness) are known as *real-time systems* [2,5].

Real-time systems are present in many different areas such as medicine, avionics, multimedia and mechatronics [13]. For some of them, when timing requirements are not accomplished, the system may not achieve the expected level of Quality of Service (QoS). This is what happens, for example, in a video transmission (multimedia application). In worst cases, missing timing requirements may have undesired consequences as for example, if we consider an automobile ABS control, in which human life may be at risk [3]. This evidenced that different applications may have different criticality levels. Indeed, for "soft" real-time systems missing deadlines may not have more serious consequences, while for "hard" real-time systems missing deadlines may cause injuries for human beings and/or environment [2, 9, 11].

Since temporal requirements play an important role for real-time systems, it is crucial to have means of guaranteeing such requirements. In fact, both *scheduling policies* and *schedulability analysis* are responsible for ensuring timeliness for such applications. To do so, system tasks are ordered according to a specific scheduling policy and a subsequent schedulability analysis is performed to assess timeliness of each task. We detail such aspects in Section 2.

Ensuring reliability is an important goal to be achieved for real-time systems. However, in terms of computational applications, the only certainty we have is that all of them may potentially fail [1]. In fact, system correctness relies on its dependability, a concept which discussed in Section 3. Also, since faults cannot be avoided and are difficult to predict [9, 11], taking such events into consideration is almost an obligation, if someone needs to guarantee QoS for real-time applications or even avoid more serious consequences.

In this paper we investigate the impact of errors in real-time applications considering a specific scheduling policy. To do so, we defined a simulation environment, presented in Section 4 and developed a simulation tool, detailed in Section 5, which aims at measuring fault resilience for a particular set of real-time systems. Last, in Section 6, some conclusions and future works are drawn.

2 Real-Time System Overview

A real-time system is a computer system in which both timing and logical requirements must be respected. Thus, the correct behavior of such a system depends not only on the integrity of produced logical results (also known as "correctness"), but also on the time at which they are produced ("timeliness") [2]. Examples of real-time systems include current control laboratory experiments, vehicle control, nuclear plants and flight control systems [13].

Usually, real-time systems are structured as a set of n periodic tasks $\Gamma = \{\tau_1, \tau_2, \ldots \tau_n\}$. A given task τ_i represents a function, routine (or subroutine) or any code snippet. Each task τ_i has attributes such as an execution cost C_i, a deadline D_i, an activation period T_i and a recovery execution cost $\overline{C_i}$. Thus, a periodic task can be described as an ordered tuple $\tau_i = (C_i, T_i, D_i, \overline{C_i})$.

Tasks are executed in a specific order called execution scale. Such a scale is defined based on some heuristics, known as *scheduling policy*. Several scheduling policies have been addressed in literature and most of them are priority oriented [3, 9, 10], which means that tasks are ordered according to its priority.

A well-known priority oriented scheduling policy is Rate Monotonic (RM), according to which tasks with shortest periods have higher priority. Clearly, this is a fixed-priority policy, since tasks period are defined offline and do not change during system execution.

After a scheduling policy is chosen for a given task set Γ is it important to assess if any task $\tau_i \in \Gamma$ may miss its deadline. To do so, we perform some tests, also known as schedulability analysis, which aims at determining if a given task set is *feasible*. In other words, such tests determine if any task $\tau_i \in \Gamma$ misses its deadline. Clearly, schedulability analysis is strongly linked with the chosen scheduling policy. In this paper we address the analysis based on Processor Utilization Analysis, which is discussed in Section 2.1.

2.1 Processor Utilization Analysis

According to this approach, the schedulability of a given task set is assessed based on processor use. Indeed, processor utilization U, for a given task set Γ composed of n a periodic and independent real-time task is given by:

$$U = \sum_{i=1}^{n} \frac{C_i}{T_i} \quad (1)$$

Regarding Rate Monotonic, if we assume a periodic task set Γ in which tasks period are equal to their deadlines, we state that Γ is schedulable if:

$$U \leq n(\sqrt[n]{2} - 1) \quad (2)$$

For RM, Processor Utilization Analysis is a sufficient schedulability test, which means that it is not able to ensure schedulability for all task sets. In fact, it has been proven that [6] if:

$$U \leq \ln 2 \quad \textbf{(3)}$$

The task set is schedulable with RM. Otherwise the analysis does not guarantee schedulability. Also, Rate Monotonic is considered an optimal algorithm for systems in which tasks period are equal to their deadlines ($T_i = D_i$) [6].

3 Fault-Tolerant Real-Time Systems

Faults are random events that cannot be predicted or avoided. Actually, the only certainty we have is that all computational applications potentially fail [1]. Indeed, a fault may be caused by several different events, as for example cosmic radiation, hardware fatigue or malfunctioning, specification and/or implementation aspects.

A system is said to *fail* when there is a transition from an expected correct behaviour to an incorrect and unexpected behaviour. In other words, a fail represents a deviation from specification. The *error* is the state that leads the system to fail and *faults* are the causes of an error, which may be physical or algorithmic [1]. Indeed, applications must provide confidence in the expected operations, a concept usually addressed as *dependability*, which is related to some attributes such as availability, reliability, safety and maintainability [1, 14].

In terms of real-time system there is a concern about fault tolerance aspects, since a fault may affect the system schedulability, or in other words, may prevent tasks to meet their deadlines. For this reason, faults are considered as a threat to dependability. Thus, techniques must be implemented to deal adequately with faults, so that applications keep their correctness even in the presence of such events [1, 14].

Faults are more commonly classified based on the persistence criterion, according to which they can be transient, intermittent or permanent. *Transient* faults occur only for a given time and then disappear. An example could be electromagnetic interference. When a transient fault occurs repeatedly it is called *intermittent*, as for example a loose contact on a connector. Both transient and intermittent faults are difficult to diagnose. Last, *permanent* fault is one that continues to exist until the faulty component is repaired, as for example a lack of connectivity between two nodes in a network [10, 14].

In this paper we investigate the effects of transient faults focusing on techniques that can be used to deal with such faults, which are based on temporal redundancy. This consists of repeating the computation in time, or in terms of scheduling can be understood as re-executing a task (see Figure 2) or executing an alternative task (see Figure 3) until the system is put on a safe state [9,11].

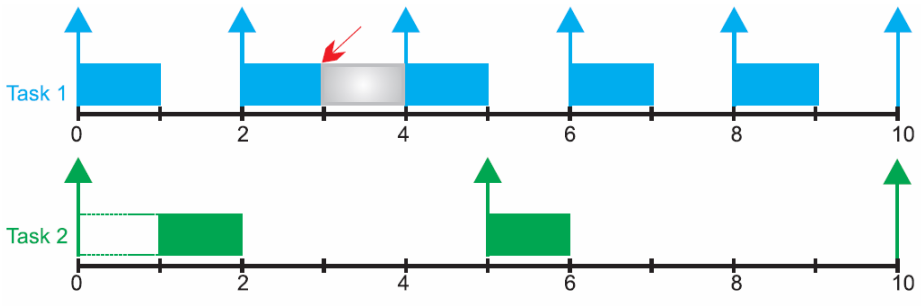


Fig. 2. Recovery based on re-execution of Task 1 under RM}

Figures 2 and 3 presents a periodic task set being scheduled, where $\Gamma = \{\tau_1 = (1,2), \tau_2 = (1,5)\}$. Observe that in Figure 2 an error occurred at $t = 3$ (red arrow), which affect Task 1. The faulty task re-executed immediately since there were no other higher priority task to execute. On the other hand, in Figure 3, an error affected Task 2 at $t = 6$ (red arrow), but it only could recover at time $t = 7$, since Task 1 was already released for execution and has a higher priority than Task 2.

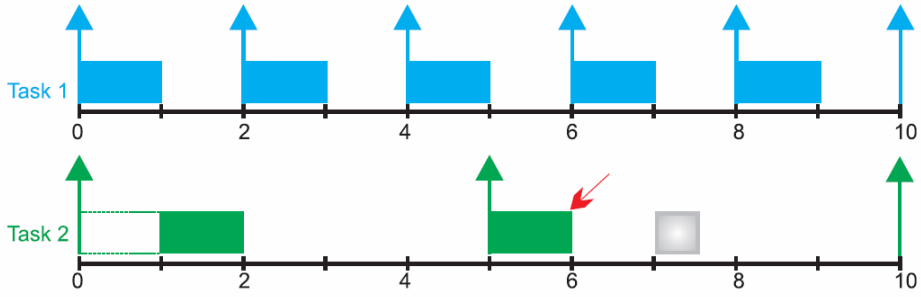


Fig. 3. Recovery based on the execution of an alternative version of Task 2 under RM.

In the following sections we present the developed tool which focus on measuring the resilience of hard real-time systems scheduled according to RM scheduling policy.

4 Simulation Environment

4.1 System Model

The assumed system model considers a task set Γ composed of n independent and periodic real-time tasks $\Gamma = \{\tau_1, \tau_2, \ldots \tau_n\}$. Each task τ_i is represented by a tuple $\tau_i = (C_i, T_i)$ where C_i is the constant worst-case execution time (wcet) of each task and T_i

is the activation period. Also, we assume that the deadline for each task is the same as its period ($T_i = D_i$).

Tasks are scheduled according to Rate Monotonic, since this algorithm deals with fixed priority tasks and is widely used for embedded critical applications. Also, schedulability analysis is performed with Processor Utilization Analysis.

4.2 Fault Model

Assuming a specific fault model is a difficult task since faults are random and cannot be predicted. We consider that the system is subject to multiple transient faults which can occur at any time instant.

Also, we represent the fault resilience of a given system through the maximum number of errors the system can handle and keep its correct behavior. To do so, we use a random function in MATLAB® to generate the number of errors that will affect each system. Also, the time instant in which errors occur is also determined through a random procedure.

We discard errors that occur at time instants in which no task is executing, since such errors will not affect system behavior. We assume that fault detection occurs implicitly, at the end of each task execution, since the focus of the work is not the detection procedure, but system behavior after recovery strategies.

4.3 Recovery Model

The recovery model describes the strategy used to put the system in a safe state. Indeed, we consider two possible actions: (i) faulty task re-execution or (ii) execution of an alternative task. Both strategies are defined offline, before running the system, and are performed in idle time instants available in execution scale.

5 Simulation Tool

The general overview of the developed tool can be seen in Figure 4. The tool was developed in MATLAB®, due to its versatility on numerical analysis, encapsulated functions and graphics.

Fig. 4. Framework of Simulation Environment

The first step to use this tool is to input a *schedulable task set*. In case the user has no previously generated task set, it is possible to generate a random one inside developed environment. To so, the user only has to choose the number of tasks to be generated. In case the tool generates the task set, it also tests if it is schedulable, through processor utilization analysis.

After, the user has to generate the *number of errors* that will affect the task set. As mentioned before, such a number is randomly generated by the tool. The user only defines a lower and upper bound, which will represent the interval in which the number of errors will be in. Based on the number of errors, the tool generates *random time instants* in which errors will occur. The screen of MATLAB® running the simulator can be seen in Figure 5.

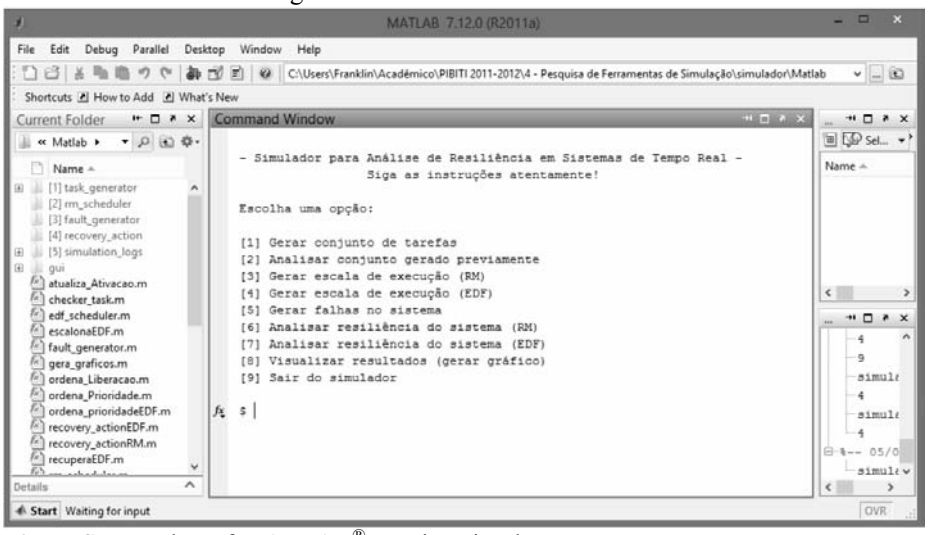


Fig. 5. Screen shot of MATLAB® running simulator.

The *simulation environment* will generate an execution scale, which takes into consideration Rate Monotonic, as scheduling policy, the defined recovery scheme (re-execution or alternative task code) and the time instants in which errors occur. Based on those values, the system resilience is defined and results can be graphically checked.

Briefly, the simulator executes the following steps, given the inputs described in Figure 4:

- Identify tasks affected by errors;
- Identify idle time after each faulty task, which can be used for recovery;
- Verify the possibility of re-executing the faulty task or executing an alternative code, respecting tasks priority (including the simultaneous verification of space for recovery and maximum execution time);
- Graphically analyze the resilience of the system, through graphically generated execution scale.
- Inform the number of errors and time instant which makes the system unschedulable.

To make things clear let us consider the following example:

Example 5.1. Assume a task set $\Gamma = \{\tau_1, \tau_2\}$ composed of two independent and periodic tasks where $C = (3, 3)$ and $D = T = (8, 12)$. Tasks are scheduled according to RM and in case of faults, tasks are re-executed. In other words, $C_i = C_i$.

Fig. 6. Execution Scale for Example 5.1.

The system is simulated during the hyperperiod $h = lcm(8, 12) = 24$[1] to ensure that all system execution will be considered.

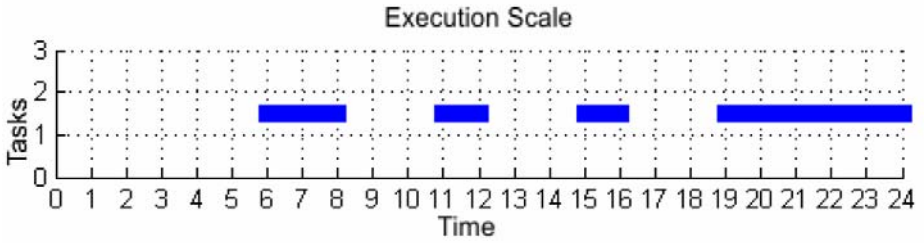


Fig. 7. Idle processor time for Example 5.1, graphically represented in tool.

The first chart presented in Figure 6 presents the execution scale for the given task set. The random number of faults that this task set is subject to is $n_f = 2$ and the random time instants in which they occur was $t_f = (3, 18)$. This is shown by a red mark in the chart. Detected errors are indicated by green circles.

Figure 7 evidences the idle processor time, which are represented in blue. Finally, Figure 8 presents the fault-tolerant real-time schedule.

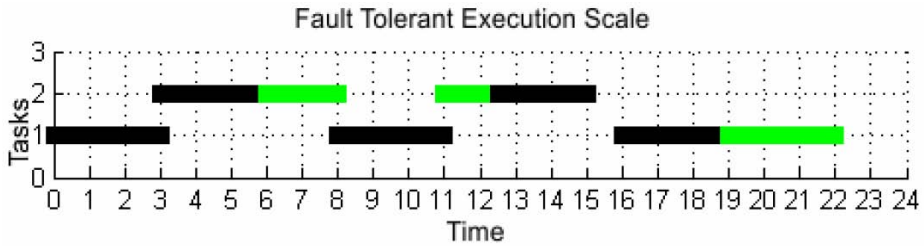


Fig. 8. Fault Tolerant Scheduling for Example 5.1 assuming errors at $t_f = (3, 18)$.

[1] *lcm(x, y)* is the function which calculates the least common multiple of input parameters ,in this case, x and y. Usually systems are simulated during the hyperperiod, since it contains all system behavior.

It is important to mention that depending on the time instant that errors occur, the system may not be schedulable, even if it is subject to the same number of errors. Observe Figure 9, which presents the same task set described in Example 5.1 subject to two errors that happens at $t_f = (2, 3)$.

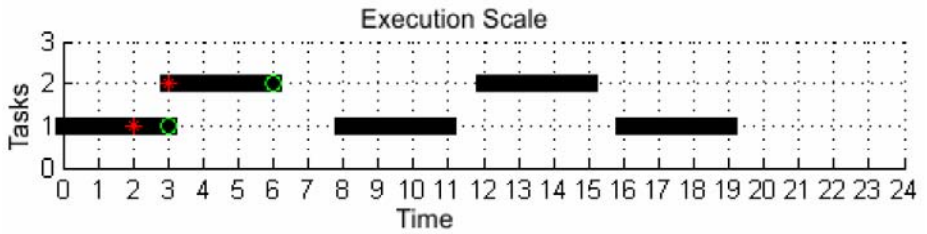


Fig. 9. Execution Scale for Example 5.1 assuming errors at $t_f = (2, 3)$.

Observe that in this case, recovery of both faulty tasks is not possible, since the available idle time (same as presented in Figure 7) is not enough for recovering tasks τ_1 and τ_2. before their respective deadlines. The graphic presented by simulator is according to Figure 10, confirming that the fault-tolerant scheduling is not feasible.

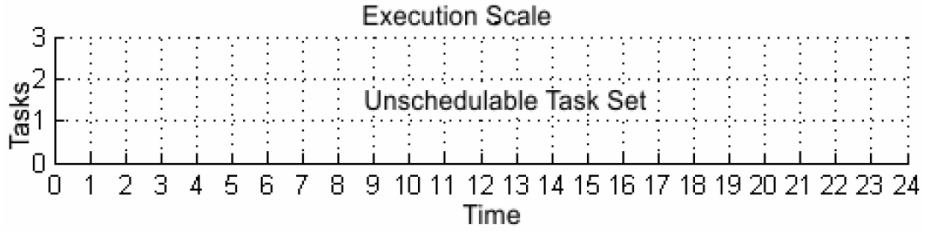


Fig. 10. Fault Tolerant Scheduling for Example 5.1 assuming errors at $t_f = (2, 3)$.

6 Conclusions and Future Work

Real-time systems have been used in a wide range area, as for example to control industrial processes. For most of these applications, missing timing requirements imply in a loss of Quality of Service or in worst cases may cause social, economic and/or environmental injuries. In this context, it is extremely necessary to deal with unpredictable and random events, such as faults, so that they interfere minimally in systems operation. In this paper we developed a simulation tool in MATLAB® environment to deal with fault-tolerant real-time scheduling, so that errors consequences can be envisioned, before system is put on operation.

One of our goals is to have an approximation between theoretical and practical models. This will enable more detailed studies and previous use of simulations before the applications are put into production. As future work we aim at simulating more robust systems, to evaluate better our preliminary results. Also, we focus on extending scheduling policies, so that EDF [6] is also considered.

7 Thanks

The authors would like to thank the Federal Institute of Bahia (IFBA) for all support and National Council for Scientific and Technological Development (CNPq) for funding this work.

References

1. Algirdas Avizienis, Jean-Claude Laprie, Carl Landwehr, and Brian Randell. Basic concepts and taxonomy of dependable and secure computing. IEEE Transactions on Dependable and Secure Computing, 1(1):11–33, 2004.
2. Jean-Marie Farines, Joni da Silva Fraga, and Rômulo Silva de Oliveira. Sistemas de Tempo Real. Departamento de Automa¸c˜ao e Sistemas - Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, 2000.
3. Sunondo Ghosh, Rami Melhem, and Daniel Moss´e. Fault-Tolerant Scheduling on a Hard Real-Time Multiprocessor System. In Proccedings of Eighth International Symposium on Parallel Processing, pages 775 – 782, April 1994.
4. Rolf Isermann. Modeling and design methodology for mechatronic systems. IEEE/ASME Transactions on Mechatronics, 1(1):16–28, 1996.
5. Li Jie, Guo Ruifeng, and Shao Zhixiang. The Research of Scheduling Algorithms in Real-Time System. In International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE'10), volume 1, pages 333 – 336, June 2010.
6. C. L. Liu and James W. Layland. Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment. Journal of the ACM, 20(1):46–61, 1973.
7. Ren C. Luo. Sensors and actuators for intelligent mechatronic systems. In 27th Annual Conference of the IEEE on Industrial Electronics Society, 2001. IECON'01, volume 3, pages 2062–2065, 2001.
8. Paulo Eigi Miyagi and Emilia Villani. Mecatrônica como solução de automação. In Revista Ciências Exatas. Universidade de Taubaté, 2004.
9. Flávia Maristela Nascimento, George Lima, and Verônica Cadena Lima. Deriving a fault resilience metric for real-time systems. In Workshop de Testes e Tolerância a Falhas, August 2009.
10. Flávia Maristela Santos Nascimento. A Simulation-Based Fault Resilience Analysis for Real-Time Systems, 2009.
11. George Lima, Flávia Maristela Santos Nascimento, Verônica Lima. Fault resilience analysis for real-time systems. In Proceedings of the 1st International Workshop on Analysis Tools and Methodologies for Embedded and Real-Time Systems, pages 35 – 38, Brussels, Belgium, October 2010.
12. Mircea R. Stan and Kevin Skadron. Power-Aware Computing. Computer, 36:35 – 38, December 2003.
13. John Stankovic. Misconceptions about real-time computing: a serious problem for next-generation systems. Computer, 21(10):10–19, 1988.
14. Andrew S. Tanenbaum and Maarten van Steen. Distributed Systems: Principles and Paradigms. Prentice Hall, 2006.

Application of Zigbee Technology for Monitoring Enviromental Variables in Greenhouses

Juan Carlos Suárez Barón[1]

[1] Assistant Research
Faculty of Engineering, Universidad Nacional Abierta y a Distancia, UNAD
Duitama, Colombia
jsuarezbaron@gmail.com

Abstract. This paper describes the application of the Zigbee standard for the development of a Wireless Sensor Networks (WSN) based system, which is used for monitoring environmental variables in greenhouses. This development allows to connect multiple wireless devices for the sake of transmitting variables such as temperature and relative humidity. The Zigbee platform was made in three stages: 1) hardware development, which includes the analysis and hardware selection; 2) construction of a network and the integration of sensors; and, 3) evaluation, in order to define the specifications of each node and scope of communication.

Keywords: Environment variables, Greenhouses, WSN, Zigbee.

1 Introduction

Greenhouses are used to reduce the influence of adverse factors that limit production and quality of crops. They include the control of environmental variables and make an efficient use of water. On the other hand, modern greenhouses covers several hundreds of square meters, where the location to measure temperature, humidity and lighting is carefully chosen in order to improve production efficiency; thus, a Wireless Sensor Network (WSN) is required.

A WSN system includes several spatially distributed devices that use sensors to monitor various conditions in several points, including temperature, sound, vibration, pressure, motion and pollutants [1]. WSN systems have been used for various applications, e.g. habitat monitoring, agriculture, industrial monitoring and control, electronics, home automation and medical health care [2]. There are different technologies for WSN; however, the technology known as Zigbee is one of most widespread. Zigbee was developed for applications where energy consumption and complexity are the main concern. Zigbee is suitable for communicating sensors, actuators and other small devices among them. It makes use of a narrow bandwidth, low energy consumption and low latency [3]. Zigbee is based on the IEEE 802.15.4 standard and defines the hardware and software described in terms of network connection, such as physical layer (PHY) layer and medium access control (MAC). Basically, the system developed in present work consists of a sensor node and a coordinator device. The sensor node is basically a data acquisition unit, and it is responsible for collecting climate variables such as temperature, relative humidity,

and light, and transmits the collected data to the coordinator station through Zigbee modules.

2 Background

Wireless sensor networks represent a significant advance over traditional invasive methods for the monitoring species, which can achieve lower costs and errors in the measurement process [4]. For instance, WSN are used for monitoring the reproductive behavior of birds in the Great Duck Island (Maine, USA), as described in [5]. This system enables biologists the analyzing of changes in the environmental conditions inside and outside the burrows during the breeding season. On the other hand, environmental conditions are also a concern. It is developed in [6] a monitoring system of the pollution caused by the emission of gases from car exhausts.
The data generated by the gas sensors are transmitted to remote stations via Zigbee modules. Similar Zigbee based systems have been used to monitor water quality in rivers and lakes, as explained in [7], [8]. In agriculture, wireless sensor networks are used to increase efficiency in the production and growth of the crop. Usually, sensed data correspond to environmental conditions such as temperature, wind speed, wind direction, soil moisture and physical and chemical properties of soil such as pH [9].
Another way to increase efficiency in crops is by water resource management; and in this respect, several systems based on sensor networks have been implemented. In [10] it is described the development of a crop irrigation control system in Pakistan. This system makes use of wireless sensor-actuator networks (WSAN) to monitor environmental parameters, which are sent through Zigbee modules to a computer. These variables serve as inputs to the control system. Additionally, the authors in [11] propose the design and implementation of an irrigation system based on low-cost Zigbee technology. Monitored variables are temperature and humidity. The other hand, in [12] it is introduced the use of a wireless sensor network based on Zigbee technology (ZWSN). The climatic variables monitored are temperature, speed and direction of air, relative humidity and rainfall. The data and images related to the amount of leaves and fruits are sent to a personal digital assistant (PDA), which processes and displays the information in order to monitor, in a detailed way, the evolution of diseases. In particular, the impact caused by fruit fly is tracked.
Finally, the authors in [13] develop a wireless sensor network based on Zigbee technology, which uses MPWiNodeZ devices, intended for precision viticulture applications. A mesh topology network is utilized to monitor the moisture content of soil, air temperature, relative humidity and solar radiation.

3 Materials and Methods

3.1 System description and hardware development

The system consists on a sensor node and a coordinator device that are communicated between them. The sensor node is basically a data acquisition unit, and it is responsible for collecting climate variables such as temperature, relative humidity and

light. In addition, the system transmits the collected data to the coordinator station through Zigbee modules. In this work, SHT71 was selected as the integrated temperature and humidity sensor chip. Regarding humidity, the operating range is from 0 to100%, and the operating range of temperature is from -40 to 125 °C. SHT71 sensors have low power consumption and fast response time.
Temperature accuracy is ± 0.4°C and the accuracy of the relative humidity is less than ±3.0 %. Therefore, SHT71 is a good solution for monitoring these variables in the agriculture field [14].
The coordinator system, who acts as a central station, is responsible for receiving the data acquired by the sensor nodes forming a star topology network. This system process, stores and provides to the user a convenient and easy way of displaying real time information by means of a GUI (Graphic user interface) on an LCD device. The functional diagram of the system is showed in Fig 1.

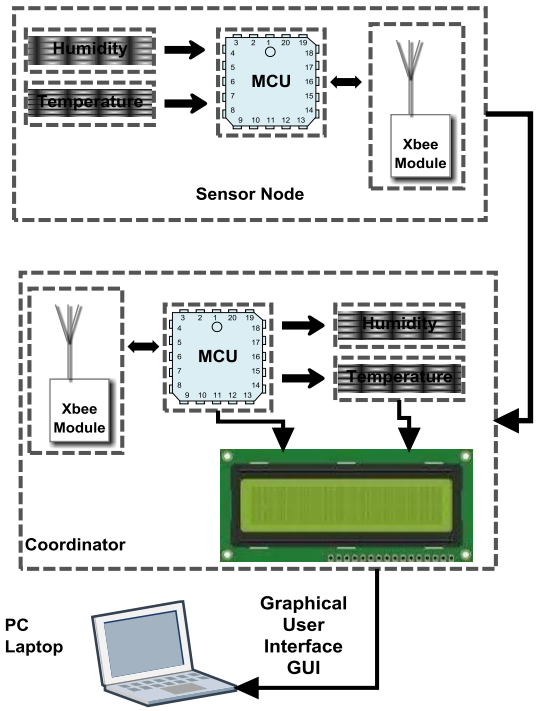


Fig. 1. Structure of wireless monitoring system. (System Overview)

3.2 Prototype network node and the integration of sensors.

The sensor node is composed by four (4) elements:
The module of sensors
The processing module
The wireless communication module

The power supply module

3.2.1 Sensor Node Design

The sensor module is responsible for collecting information about temperature and relative humidity. The processing module stores and processes data collected by the sensors, and controls the operation of the sensors node; which are achieved by using a microcontroller (MCU). An HCS08 based MCU, the MC9S08JM16, was selected as the main control chip of the sensor node. HCS08 MCUs are suitable due to their processing and memory capacities, being sufficient to support Zigbee. The wireless communication module communicates with other nodes, allowing the exchanging (receiving/transmitting) of data. The power supply device provides energy to the sensor, processing and the wireless communication modules. The power supply of the sensor node corresponds to a 9V alkaline (Zn/MnO2) battery. Using a battery is a low cost, portable and low maintenance solution. On the other hand, the Xbee module, SHT71 sensor and MC9S08JM60 microcontroller require 3.3 V, which are provided by an LM1117 regulator. The block diagram of the sensor node or end device shown in Fig 2.

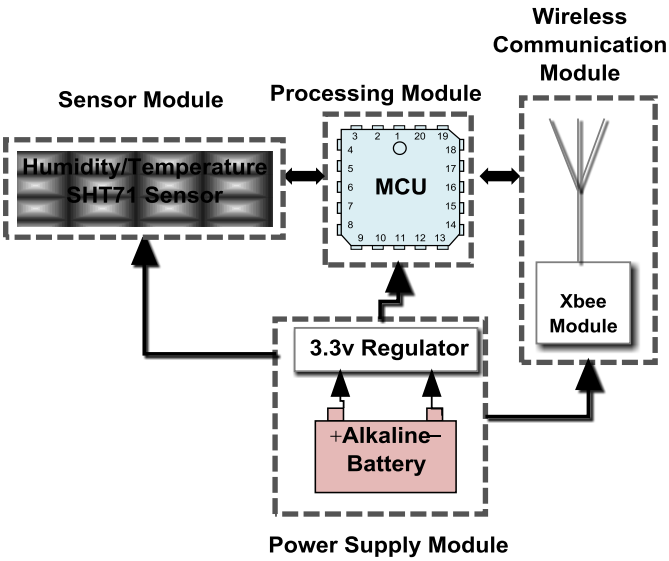


Fig. 2. Block diagram of sensor node.

The final design consists of two PCB layers. The top layer is used to place the XBee module, LEDs (power, Rx and Tx indicators), microcontroller, SHT71 sensor, battery plug, and on/off switch. Bottom layer is used to place3.3V voltage regulator and as a ground plane under XBee module to minimize any interference cause by the RF signals. An important aspect of the design was miniaturization. Therefore most circuitry components used for the sensor station are either surface-mounted (SMD) or are very small in size. The final design of the sensor node is shown on Fig 3.

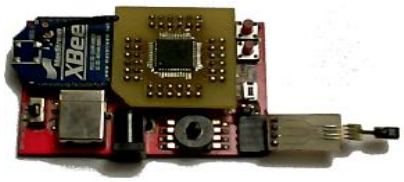


Fig. 3. Final design of the sensor node.

3.2.2 Coordinator Design

The coordinator receives the signals from the sensor node, and then it integrates and stores the data automatically. The coordinator is composed by five parts: processing module, wireless communication module, power module, display module, and USB communication module. The processing module controls the operation of the sensor nodes; and, stores and processes the collected data. The wireless communication module is responsible of receiving/transmitting data from/to the sensor node. The power supply module provides power to the other modules. The data logger is responsible for storing the sensor data, which are displayed on a liquid crystal display LCD. The union of these allows the coordinator to periodically receive data from the sensor node. The block diagram of coordinator system is shown in Fig 4.

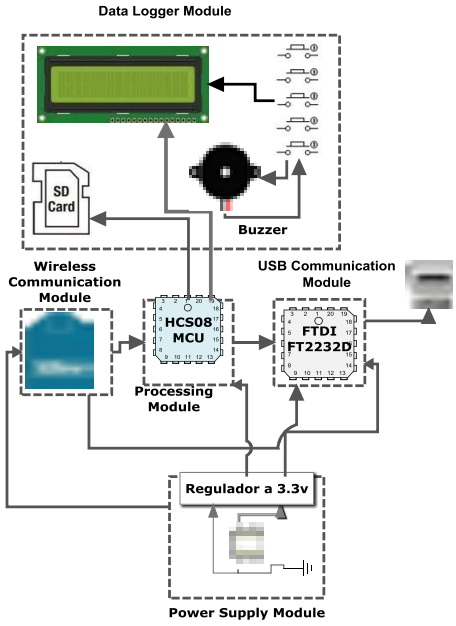


Fig. 4. Block diagram of coordinator device.

For the coordinator, as well as in the sensor node, the XBee modules based on the IEEE 802.15.4/Zigbee Wireless Personal Area Network (WPAN) standards to build a

low-power, low-maintenance, and self-organizing WSN [15] was used. Small size, low power, low cost and long battery life are the reasons of using ZigBee.

In case of the hub, the power supply is derived from the PC's USB port via a connector type B. The data logger module includes an LCD and five buttons that are utilized select the physical variables to be displayed and stored in an SD memory card. In order communicate the PC to the coordinator, it was used an FT2232D FTDI chip. The final design of the sensor coordinate is shown on Fig 5.

Fig. 5. Final design of the coordinator.

3.3 Sensor Configuration

The temperature sensor output was set to 12 bit format, and RH was configured to 8 bit format; thus, it is achieved an accuracy of 0.04 C and 0.5 for temperature and RH variables, respectively. The sensor and the microcontroller interact by using I²C protocol, hence only two pins on the microcontroller are required. One of the pins will be used for synchronization while the other is utilized for bidirectional data transfer between the two devices. The pin on the microcontroller that is used for Rx/Tx was pulled up in order to prevent signal contention. In order to interact adequately with the sensor, a specific sequence of events must be followed. The flow chart in Figure 8 illustrates the events to be followed in order to request the sensor to take measurements such that the information can be read.

3.4 Communication between devices

The communication of the SHT71 sensor with the microcontroller MC9S08JM16 was through I²C module. This is possible because there is not any device connected to the I²C output of the microcontroller, thus it does not generate interferences [16]. Communication with this sensor requires the implementation of a protocol, which is very similar but not compatible with the I2C standard. Therefore, the context integration is strictly controlled. The protocol includes a start condition, and data block both reading and writing with ACK bit. The communication is based into two

pins; a clock (SCK), that is used to synchronize the microcontroller and the sensor; and, the bidirectional data pin.

4. Results and Discussion

After verifying proper communication between the various elements of the system built, we proceeded to test the proper connectivity to all the prototype and the proper functioning of the tasks. In order to emulate a greenhouse environment, it was set a space containing two ornamental plants. Near each plant, it was located a sensor node that measures relative humidity and temperature around the plants. The sensor node outcomes are visualized on the LCD. Fig 6 shows the plants and the corresponding measures shown on the LCD.

Fig. 6. Plants 1y 2 of the test displayed in the LCD.

The results of test are depicted in Fig 7 and Fig 8. The results obtained from the experiments show small variations between the readings of the SHT71 sensor for the two tests.
Future experiments will entail comparing data collected from these sensors with calibrated standard devices for the sake of obtaining more accurate results.

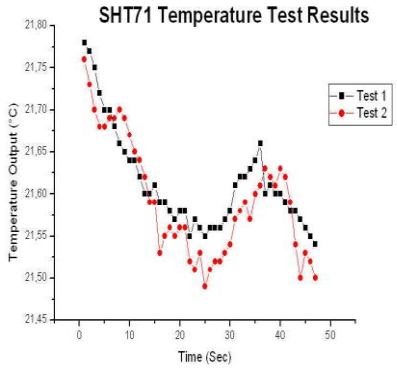


Fig. 7. SHT71 experimental results (Ambient Temperature)

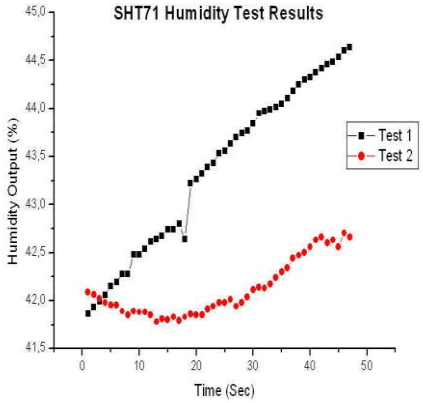


Fig. 8. SHT71 experimental results (Relative Humidity)

Other experiment consisted on verifying the communication between the two Xbee modules. Figure 9 describes the distances that separate the sender (red icon) of the receptor (blue icon). Measures were taken of packet reception and level RSSI (Received Signal Strengh Indicator) of the received packet to 30 m, 60 m and 100 m away. For the test used two Xbee devices, two laptops and two USB boards for development of digi connected. In the X-CTU software was used Range-test option with its default settings. This configuration is sending 32 bytes of data from one device to another, which returns the data frame to the origin. Others experimental results were based on the Lost Packets with values between 0 to 1000 and LQI within typical values between -95dBm to -18dBm according to IEEE 802.15.4 [16]. Below are the results of the experimental measures of three sequences, with 5, 10 and 20 bytes of payload size per packet, and a Zigbee Network implementation with Monitoring Environmental Variables network. In Fig 10, measures of LQI and Lost packages are shown respectively for 5 bytes in payload per packet.

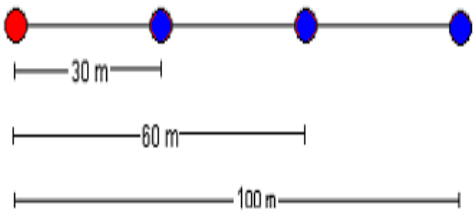


Fig. 9. Diagram for test of outdoor.

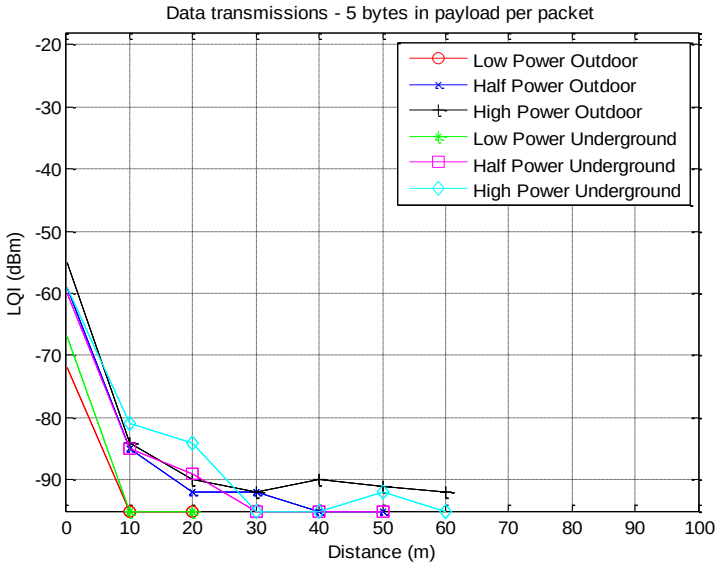


Fig. 10. Graphics of LQI vs. distance between devices

5. Conclusions and Future Work

In this work, it is presented a wireless solution for greenhouse monitoring. The system is based on HCS08 Freescale MCU´s, which monitors environmental variables through SHT71 humidity/temperaure sensor and uses and Xbee modules. Also, it is shown the design of the wireless nodes, network establishment and the software system. Monitoring system is based on ZigBee standard and provides nearly unlimited installation of transducers, which increases network robustness and reduces considerably
installation costs. The designed wireless monitoring system uses different sensors and has capability to measure different types of environmental parameters.
Developed system helps farmers to increase the harvest production with a better quality. Additionally, it has capability to detect changes in the environment. Finally, the system was tested in a greenhouse located in Boyacá-Colombia. It is concluded that the ZigBee-based monitoring system is a good solution for greenhouse monitoring. As future work, I aim to develop a neural network system in order to

carry out intelligent control; this element will be constructed together with a model for data mining and system decision support.

References

1. Aakvaag, N., Frey, J.-E.: Redes de sensores inalámbricos. Revista ABB, 39-42 (2006)
2. KeKeshtgari, M., Deljoo, A.: A Wireless Sensor Network Solution for Precision Agriculture Based on ZigBee Technology. Journal Wireless Sensor Network IV, 25-30 (2012)
3. Zigbee Alliance: Zigbee Specification 053474r17. Available at: http://www.zigbee.org
4. Cifuentes García, C.: Diseño e implementación de una red inalámbrica de sensores aplicados a la instrumentación biomédica. Tésis de maestría. Universidad Nacional de Entre Ríos, Oro Verde (2010)
5. Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D., Anderson, J.: Wireless Sensor Networks for Habitat Monitoring. In : WSNA'02 (2002)
6. Eren, H., Al-Ghamdi, A., Luo, J.: Application of ZigBee for Pollution Monitoring Caused by Automobile Exhaust Gases. In IEEE, ed. : SAS 2009 IEEE Sensors Applications Symposium, New Orleans, LA (2009)
7. Wang, X., Ma, L., Yang, H.: Online Water Monitoring System Based on ZigBee and GPRS. Science Direct. Procedia Engineering(15), 2680 – 2684 (2011)
8. Azwan Nasirudin, M., Nurulhaiza Za'bah, U., Sidek, O.: Fresh Water Real-Time Monitoring System Based on Wireless Sensor Network and GSM. In IEEE, ed. : 2011 IEEE Conference on Open Systems (ICOS2011), Langkawi, Malasia, pp.354-357 (2011)
9. Kalaivani, T., Allirani, A., Priya, A.: A survey on Zigbee Based Wireless Sensor Networks in agriculture. In IEEE, ed. : 3rd International Conference on Trendz in Information Sciences and Computing (TISC) 2011, pp.85-89 (2011)
10. Aqeel-ur-Rehman, S., Yousuf, H., Nawaz, F., Kirmani, M., Kiran, S.: Crop irrigation control using Wireless Sensor and Actuator Network (WSAN). In IEEE, ed. : International Conference on Information and Emerging Technologies (ICIET) 2010, pp.1-5 (2010)
11. Zhou, Y., Yang, X., Wang, L., Ying, Y.: A wireless design of low-cost irrigation system using ZigBee technology. In Society, I., ed. : 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing , pp.572 - 575 (2009)
12. Jiménez, A., Ravelo, D., Gómez, J.: Sistema de adquisición, almacenamiento y análisis de información fenológica para el manejo de plagas y enfermedades de un duraznero mediante tecnologías de agrícultura de precisión. Tecnura. Redalyc XIV(27), 41-51 (2009)
13. Morais, R., Fernandes, M., Matos, S., Serôdio, C., Ferreira, P., Reis, M.: A ZigBe multi-powered wireless acquisition device for remote sensing applications in precision viticulture. E. B.V II(62), 94-106 (2008)
14. Kovács, Z., Marosy, G., Gyula, H.: Case study of a simple, low power WSN implementation for forest monitoring. In IEEE, ed. : Biennial Baltic Electronics Conference (BEC2010) , Tallinn, Estonia, pp.161-164 (2010)

15. Digi International Inc: XBee ZNet2.5/XBee-PRO ZNet2.5 OEM RF Modules, Product Manual v1.x.4x - ZigBee Protocol For OEM RF Module Part Numbers: XB24-BxIT-00x. In: Digi International Inc 11001 Bren Road East Minnetonka, MN 55343877 912-3444 or 952 912-3444. Available at: http://www.digi.com

16. Sensirion AG: Datasheet SHT7x (SHT71, SHT75) Humidity and Temperature Sensor IC. (Accessed 2011) Available at: http://www.sensirion.com/fileadmin/user_upload/customers/sensirion/Dokumente/Humidity/Sensirion_Humidity_SHT7x_Datasheet_V5.pdf

Inversión de prioridades: prueba de concepto y análisis de soluciones

Raúl Benencia, Luciano Iglesias, Fernando Romero y Fernando G. Tinetti**

Instituto de Investigación en Informática III-LIDI
Facultad de Informática, UNLP
rbenencia@linti.unlp.edu.ar,li@info.unlp.edu.ar,
{fromero,fernando}@lidi.info.unlp.edu.ar,
http://weblidi.info.unlp.edu.ar

Resumen La planificación de tareas es el punto crucial de un sistema de tiempo real. Dicha función es llevada a cabo por el planificador del sistema operativo, diseñado para poder cumplir con las restricciones temporales de dichas tareas, teniendo en cuenta sus valores de prioridad. Al haber recursos compartidos por estas tareas, se produce el efecto llamado inversión de prioridades. En este trabajo se analiza dicho efecto y se evalúan las soluciones implementadas para este problema en el Sistema Operativo de Tiempo Real GNU/Linux con parche RT-PREEMPT.

Keywords: Planificación de tareas de tiempo real, inversión de prioridades, GNU/Linux con parche RT-PREEMPT, Sistemas Operativos de Tiempo Real

1. Introducción

Los sistemas de tiempo real requieren ser correctos lógica y temporalmente. Se deben respetar las restricciones de tiempo en la ejecución de las tareas. De acuerdo a la función que realizan, las tareas pueden requerir plazos estrictos o plazos más relajados. Esto conlleva a la necesidad de una planificación basada en prioridades fijas, de manera de asegurar los límites estrictos. Este tipo de planificación es respecto de uno de los recursos compartidos por las diversas tareas, implicadas en la ejecución, la CPU. De tal manera que si un proceso que implementa una tarea de baja prioridad está usando la CPU y otro proceso con una prioridad mayor requiere su uso, está previsto el desalojo del proceso de menor prioridad para asignársela al de mayor prioridad. El resto de los recursos compartidos por las tareas suele planificarse por demanda (FCFS). Al producirse situaciones de bloqueo de estos recursos, que requieren usarse en forma exclusiva (región crítica), puede pasar que el proceso de mayor prioridad quede relegado por uno de menor prioridad. Esta situación se denomina *inversión de prioridades* [2] y se puede agravar si procesos de prioridad intermedia logran hacerse de la CPU antes de que la tarea de prioridad baja libere el recurso que espera la de

** Investigador CICPBA

alta prioridad. Con lo cual la duración de la inversión de prioridades puede ser ilimitada y no permitir a la tarea de mayor prioridad cumplir con sus restricciones temporales. Los Sistemas Operativos de Tiempo Real suelen contar con mecanismos para atenuar este problema. En rigor, estos mecanismos no evitan la inversión de prioridades, solo la inversión de prioridades ilimitada.

2. Descripción del problema y soluciones existentes

Los Sistemas de Tiempo Real suelen realizar tareas con diferente nivel de urgencia. Si bien todas las tareas pueden tener restricciones temporales para su ejecución, estas restricciones pueden ser estrictas (hard real-time) o menos estrictas (soft real-time) [3]. Esto es manejado por los Sistemas Operativos de Tiempo Real de diferentes maneras. Esto se conoce como planificación de tareas. Uno de los métodos de planificación de tareas se realiza asignando diferentes niveles de prioridad, de tal manera que si se debe incumplir una restricción temporal sea de las tareas con menor nivel de prioridad. A su vez, estas tareas suelen estar implementadas por threads, que se caracterizan por compartir memoria y otros recursos. Esto genera un problema llamado inversión de prioridades, que se describe a continuación: considérese un proceso con prioridad alta, el proceso H, y otro con prioridad baja, el proceso L. Ambos procesos interactúan con el recurso r y, para evitar inconsistencias, deben proteger sus respectivas secciones críticas con un mutex, por ejemplo. Si el proceso H intenta utilizar r mientras L mantiene un lock sobre el mismo, entonces H no podrá continuar su tarea hasta que L no libere el recurso. Hasta aquí se plantea una situación normal. El procedimiento correcto a seguir es asignar el procesador a L para que libere a r lo más pronto posible, para que luego H pueda adquirir el lock sobre el recurso y así continuar su tarea. Sin embargo, el problema de la inversión de prioridades se presenta cuando H espera a que L libere el lock, y mientras L intenta liberar el recurso, el mismo es interrumpido por un proceso de prioridad superior a L pero inferior a H. De esta forma, tanto el proceso L como el proceso H se ven rezagados por el proceso de prioridad media M. En algunos sistemas la inversión de prioridades puede pasar desapercibida puesto que, a pesar de las demoras, las restricciones de tiempo se cumplen y por lo tanto el sistema de tiempo real no falla. Sin embargo, existen numerosas situaciones donde la inversión de prioridades puede causar problemas críticos. Si un proceso de prioridad alta entra en estado de inanición de los recursos que precisa, puede provocar una falla en el sistema que active medidas correctivas, como un watch-dog que reinicie por completo todo el sistema.

2.1. El problema en un caso real: El Rover enviado a Marte

Tal vez el caso real más conocido de inversión de prioridades fue el que ocurrió con el rover que se envió al planeta Marte en la misión Mars Pathfinder [4] [5]. La misión Mars Pathfinder fue catalogada como perfecta a los pocos días de su aterrizaje en la superficie marciana, el 4 de julio del año 1997. Durante

varios días el rover envió cantidades voluminosas de datos, tales como imágenes panorámicas. Sin embargo, luego de pocos días de cumplir con las solicitudes requeridas desde el planeta Tierra, y no mucho después de que el rover comenzara a recolectar datos meteorológicos, el sistema operativo del robot comenzó a reiniciarse continuamente ocasionando severas pérdidas de datos. El problema se encontraba en la administración de las prioridades de VxWorks, el kernel de tiempo real embebido que usaba el Pathfinder. El planificador de VxWorks utilizaba apropiación por prioridades entre los threads. Las tareas en el rover eran ejecutadas como threads. La asignación de prioridades a los mismos se calculaba reflejando la urgencia relativa de dichas tareas. Además, el Pathfinder contenía un bus de información, similar a un área de memoria compartida, utilizado para comunicar información entre los distintos componentes de la nave. El kernel VxWorks proveía una tarea de alta prioridad dedicada a la administración de este bus, cuya función era mover ciertos tipos de datos desde y hacia el mencionado canal. La recolección de datos meteorológicos se realizaba con poca frecuencia en un thread de baja prioridad, y los datos adquiridos se distribuían utilizando el bus de información. Cuando los datos se distribuían por el bus, se adquiría el lock sobre un mutex, se escribía sobre el bus, y se liberaba el lock. Si una interrupción causaba la apropiación de la CPU mientras el thread de baja prioridad mantenía el bloqueo sobre el bus, y el thread de alta prioridad que administraba dicho bus intentaba adquirir el mismo mutex con el objetivo de recibir estos datos, entonces dicho thread se bloqueaba hasta que el thread de baja prioridad liberase el mencionado mutex. Finalmente, el rover también contenía una tarea dedicada a la comunicación que corría sobre un thread con prioridad media. La mayor parte del tiempo esta combinación de threads funcionaba correctamente. Sin embargo, con muy poca frecuencia ocurría que el thread de comunicaciones, de prioridad media, era interrumpido durante un corto intervalo de tiempo cuando el thread dedicado a la administración del bus, de prioridad alta, era bloqueado para esperar por los datos meteorológicos provistos por el thread de prioridad baja. Cuando se daba esta situación, el thread dedicado a las comunicaciones impedía que el thread de baja prioridad pueda ejecutarse para liberar el mutex. Consecuentemente, el thread de administración del bus era efectivamente bloqueado por una tarea de menor prioridad, provocando de esta forma la inversión de prioridades. Luego de un tiempo prudencial, un watch-dog timer del rover notaba el desperfecto y concluía que algo había dejado de funcionar correctamente, provocando un reinicio total del sistema operativo [6].

2.2. Soluciones existentes

Se han propuesto diversas soluciones (protocolos) para el problema de la inversión de prioridades, acá se mencionan las principales [7]:

1. Herencia de prioridad (Priority inheritance)
2. Techo de prioridad (Priority ceiling)
3. Techo de prioridad inmediato (Inmediate priority ceiling)
4. Enmascaramiento de interrupciones

5. Incremento aleatorio
6. Basado en la restauración del recurso (Shadowing)

Herencia de prioridad La solución denominada herencia de prioridad propone eliminar la inversión de prioridades elevando la prioridad de un proceso con un lock en un recurso compartido al máximo de las prioridades de los procesos que estén esperando dicho recurso. La idea de este protocolo consiste en que cuando un proceso bloquea indirectamente a procesos de más alta prioridad, la prioridad original es ignorada y ejecuta la sección crítica correspondiente con la prioridad más alta de los procesos que está bloqueando. Por ejemplo, vuélvanse a considerar los procesos L, M y H de prioridad baja, media y alta respectivamente. Suponer que H está bloqueado esperando a que L libere un lock sobre un recurso compartido. El protocolo de herencia de prioridad, entonces, requiere que L ejecute su sección crítica con la prioridad de H. De esta forma, M no podrá apropiarse del procesador cuando L lo tenga en uso. Por lo tanto M, que es un proceso con más prioridad que L, deberá esperar a que L ejecute su sección crítica, ya que L hereda su prioridad de H. Luego, cuando L termina de ejecutar su sección crítica, su prioridad vuelve a la normalidad y el proceso H es despertado. H, que tiene mayor prioridad que M, se apropia del procesador y se ejecuta hasta terminar. Finalmente, cuando H termina su ejecución prosigue el proceso M, y por último L termina su ejecución. El kernel Linux implementa la herencia de prioridades mediante un sencillo mecanismo que se basa en prohibir la apropiación del procesador mientras se esté ejecutando código del kernel protegido por un spinlock. Las desventajas que presenta este método son [8]:

1. Este método puede causar más cambios de contexto que el de techo de prioridad
2. El anidamiento de secciones críticas protegidas por herencia de prioridad puede producir grandes demoras debido a que cada vez que se cambia una prioridad de una tarea debe ejecutarse el planificador
3. El método falla si se mezclan tareas con y sin herencia de prioridad
4. El peor caso de herencia es peor que otras soluciones al problema
5. Según algunos autores, la mayoría de las implementaciones de herencia de prioridad tienden a complicar el código de las secciones críticas, reduciendo finalmente el desempeño del sistema [9]

Techo de prioridad La solución denominada techo de prioridad es un protocolo que elimina la inversión de prioridades mediante la asignación predefinida de un techo de prioridad a cada recurso. Cuando un proceso adquiere un recurso compartido, la prioridad de dicho proceso se eleva temporalmente al techo de prioridad del mencionado recurso. El techo de prioridad debe ser más alto que la prioridad de todos los procesos que puedan acceder al recurso compartido. De esta forma, cuando un proceso se esté ejecutando con el techo de prioridad de un recurso, el procesador no podrá ser apropiado por otro proceso que quiera acceder al mismo recurso, puesto que todos tendrán menor prioridad [10]. Las desventajas que tiene este método son:

1. Se debe realizar un análisis estático de la aplicación para determinar cuáles serán los techos de prioridad de cada recurso compartido. Para realizar este análisis, todas las tareas que accedan a recursos compartidos deben conocerse de antemano. Esto puede ser difícil, o incluso imposible de determinar en una aplicación compleja [10]
2. La prioridad de los procesos aumenta y disminuye cada vez que se accede a un recurso compartido, aún cuando no haya procesos compitiendo por el mismo
3. Puede dar un bloqueo falso de threads [11]

Esta solución fue propuesta por primera vez en 1980, en uno de los primeros papers que describió el problema de la inversión de prioridades [12].

Techo de prioridad inmediato Es un derivado del Protocolo de Techo de Prioridad. En este protocolo, la tarea que accede a un recurso hereda inmediatamente el techo de prioridad del recurso. Este protocolo es mas fácil de implementar y es más eficiente (hay menos cambios de contexto) [13].

Enmascaramiento de interrupciones El enmascaramiento de interrupciones también se puede utilizar para evitar la inversión de prioridades. En este caso, las interrupciones se enmascaran cuando un proceso entra en una sección crítica, y se vuelven a habilitar cuando sale de la misma. De esta forma, la inversión de prioridades no puede ocurrir puesto que todas las secciones críticas se ejecutan sin ser interrumpidas por procesos de mayor prioridad. Para que este mecanismo funcione correctamente, todas las interrupciones deben estar deshabilitadas. Si sólo algunas se enmascaran, entonces la inversión de prioridades podrá ser re-introducida por el mecanismo de gestión de interrupciones del hardware subyacente. Esta solución al problema de inversión de prioridad suele ser encontrada en sistemas embebidos, debido a su confiabilidad, bajo consumo de recursos y sencillez en su implementación.

Las desventajas de este método son:

1. Requiere que las secciones críticas sean escasas y cortas, puesto que todo el sistema se ve bloqueado mientras un proceso se encuentra en una de ellas.
2. Mientras las interrupciones estén completamente enmascaradas los fallos de página no podrán ser atendidos [12]. Por este motivo, se desaconseja la implementación de esta solución en sistemas de propósito general.

Incremento aleatorio La solución denominada incremento aleatorio propone eliminar la inversión de prioridades mediante el incremento de la prioridad de los procesos de baja prioridad que contengan locks sobre recursos compartidos. La elección del proceso cuya prioridad será incrementada se realiza de forma aleatoria, de ahí el nombre de la técnica. El incremento de la prioridad se mantiene hasta que el lock sea liberado. Esta técnica es utilizada en sistemas operativos Microsoft Windows [14]

Basado en la restauración del recurso Se basa en la técnica de desalojo de la tarea de menor prioridad que toma el recurso al arribar una tarea de mayor prioridad. Hay dos variantes: mantener un log de lo actuado por el de menor prioridad en el recurso y deshacer los cambios de manera inversa o que la tarea de menor prioridad actúe sobre una copia del recurso, que reemplaza al recurso si la tarea de baja prioridad finaliza antes del arribo de la tarea de mayor prioridad. En caso contrario, la tarea de menor prioridad es desalojada, y el recurso aparece inalterado, como si no se hubiera ejecutado la tarea de menor prioridad. Las demoras las sufre la tarea de menor prioridad en este método.

Cabe aclarar que los sistemas operativos de tiempo real, implementan solo algunos de estos métodos:

- *FreeRTOS*: es un mini kernel de tiempo real diseñado para sistemas embebidos, preparado para funcionar en diferentes plataformas de microcontroladores. Implementa el protocolo de herencia de prioridad [15].
- *MarteOS*: es un sistema mínimo de tiempo real, basado en ADA, desarrollado por el Grupo de Computadoras y Tiempo Real de la Universidad de Cantabria. Implementa herencia de prioridad y techo de prioridad [16].
- *QNX*: es un micro-kernel de tiempo real de alto desempeño [17]. Implementa la herencia de prioridad.
- *RTAI*: Es un sistema basado en Linux que le agrega funcionalidad de tiempo Real. Este sistema es desarrollado por el Politecnico di Milano - Dipartimento di Ingeneria Aerospaziale (DIAPM). Implementa Herencia de Prioridad [18].
- *RTLinux*: las prioridades de las tareas son estáticas manejadas con dos variantes de algoritmo: FIFO o Round Robin [19]. Si bien su creador, Victor Yodaiken dice que RTLinux no soporta la herencia de prioridad por la simple razón de que es incompatible con cualquier sistema de tiempo real confiable [20] existe una extension que da la posibilidad de usar el mecanismo de herencia de prioridad y el techo de prioridad.
- *VxWorks*: de la empresa WindRiverSystem, implementa la herencia de prioridad [21].
- Linux con parche RT-preempt: implementa herencia de prioridad y techo de prioridad.

3. GNU/Linux con parche RT-PREEMPT

GNU/Linux con parche RT-PREEMPT [1] es una modificación realizada al kernel Linux que lo convierte prácticamente en apropiativo, con la excepción de algunas pocas regiones de código pequeñas. Esta modificación permite la ejecución de aplicaciones consideradas *hard real-time*.

3.1. Caso de prueba

Como parte del presente trabajo se desarrolló una aplicación para recrear el fenómeno de la inversión de prioridad en GNU/Linux con parche RT-PREEMPT.

Tiene como objetivo evaluar las soluciones de herencia de prioridad y de techo de prioridad implementadas en el kernel. El desarrollo fue realizado en el lenguaje de programación C utilizando semáforos. La aplicación ejecuta tres tareas (*threads*) de las cuales dos precisan hacer uso de una región crítica en común. Dentro de la región crítica los distintos threads hacen simplemente unas instrucciones que consumen CPU. Cada thread tiene asignada una prioridad, baja (L), media (M) y alta (H). H y L, son los threads que requieren acceder a la región crítica compartida. La aplicación se puede ejecutar de tres maneras diferentes según el algoritmo que utiliza para evitar la inversión de prioridades:

- *inversion*: en este modo el programa se ejecuta sin prevenir la inversión de prioridades.
- *inheritance*: en este modo el programa evita la inversión de prioridades utilizando el protocolo de herencia de prioridad.
- *ceiling*: en este modo el programa evita la inversión de prioridades utilizando el protocolo de techo de prioridad.

Sincronización La inversión de prioridades se manifiesta cuando los threads del programa se ejecutan sobre un sistema con un único núcleo. Debido a que este caso de prueba es una simulación, los threads se deben sincronizar para que deliberadamente suceda la inversión de prioridades. Dicha sincronización se lleva a cabo usando tres semáforos a modo de barrera, con el fin de ordenar ejecución y la solicitud de acceso al recurso compartido de los mismos. La sincronización sucede de la siguiente forma, una vez que todos los threads fueron lanzados:

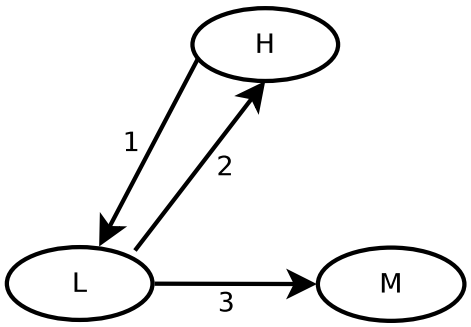


Figura 1. Se encuentran los tres threads representados por cada uno de los elipses y las flechas indican el orden y dirección de los avisos para reproducir el fenómeno de la inversión de prioridad.

1. El thread L espera la señal de que H para comenzar sus acciones. Luego de enviar la señal, H se bloquea esperando a que L bloquee el recurso compartido.

2. Luego de recibir la señal, L procede a bloquear el recurso compartido. Luego de bloquear el recurso compartido, L envía una señal a H informando este último evento.
3. Cuando H recibe la señal de que el recurso ya está bloqueado, procede a enviar una señal a M. Luego de enviar la correspondiente señal, H procede a intentar adquirir el recurso compartido.

4. Resultados

4.1. Tiempo de ejecución de los threads

La aplicación realizada para observar el fenómeno de la inversión de prioridad se ejecutó sobre un kernel Linux 3.2 con el parche PREEMPT RT [1] con un único procesador. La ejecución de *inversion* provocó la ejecución de los threads en el siguiente orden:

- H, hasta la solicitud del recurso compartido.
- M, hasta su finalización.
- L, hasta que liberó el recurso compartido.
- H, desde que se hizo del recurso compartido hasta el final.
- L, hasta su finalización.

Efectivamente se produjo la inversión de prioridades.

La ejecución de *inheritance* y de *ceiling* provocó la ejecución de los threads en el siguiente orden:

- H, hasta la solicitud del recurso compartido.
- L, hasta que liberó el recurso compartido.
- H, desde que se hizo del recurso compartido hasta el final.
- M, hasta su finalización.
- L, hasta su finalización.

En la tabla 1 pueden apreciarse los tiempos de ejecución de cada uno de los threads. Las entradas indicadas en la tabla no consideran el tiempo transcurrido entre que el thread L libera el recurso y termina su ejecución ya que no es de interés para observar el desempeño general del sistema. Como ya se mencionó estos dos mecanismos (herencia y techo de prioridades) no permiten garantizar las restricciones temporales que eventualmente podría tener la tarea de alta prioridad (H), ya que queda atada a la liberación del recurso compartido por parte de la tarea de baja prioridad (L), pero si evitan que tareas de una prioridad intermedia como M, se ejecuten antes que H.

4.2. Sobrecarga del Sistema Operativo

La implementación de cada una de la soluciones sobrecarga al sistema operativo con tareas que debe realizar al momento de adquirir o liberar un recurso.

	Inversion	Inheritance	Ceiling
L	6368	3164	3121
M	3199	9594	9550
H	9559	6400	6340

Cuadro 1. Tiempos de ejecución de cada uno de los threads, expresados en milisegundos.

Por este motivo, no utilizar ninguno de los protocolos que resuelve el problema sería como un caso base sin sobrecarga. La solución de herencia de prioridad en cada solicitud de bloqueo de un recurso compartido, debe analizar si el recurso está o no bloqueado por otro proceso de menor prioridad y de ser así pasarle la prioridad a este para que pueda finalizar lo antes posible. De la misma forma cuando libera el recurso debe analizar si tiene una prioridad heredada para restablecer su prioridad original. Por otro lado, la solución de techo de prioridades debe en cada solicitud de bloqueo determinar la prioridad techo entre todos los procesos que están a la espera del recurso y asignárselo al proceso que tiene el recurso para que este se ejecute inmediatamente. En el momento de la liberación del recurso debe asignar la prioridad techo al proceso de mayor prioridad de entre los que esperan por dicho recurso.

Se realizó en GNU/Linux con parche RT-PREEMPT un bloqueo de un semáforo[1]. Dicho bloqueo se realizó con la seguridad de que el semáforo no estaba bloqueado previamente, con la intención de hacer un primer análisis de la sobrecarga que cada algoritmo que soluciona la inversión de prioridad conlleva. Con respecto a la opción no impedir la inversión de prioridad, el bloqueo del semáforo con herencia de prioridad llevó un 47 % más de tiempo, mientras que usando el techo de prioridad llevó un 1555 % más de tiempo.

5. Conclusiones y líneas futuras

Con respecto a la ejecución de la prueba de concepto, mencionada en 4.1, y como ya se mencionó allí ni la solución por herencia de prioridades ni la solución de techo de prioridades garantizan el determinismo necesario en un sistema *hard real-time* ya que están ligados a la liberación del recurso compartido por parte de la tarea de baja prioridad. La solución de restauración del recurso si puede garantizar el determinismo necesario, penalizando a las tareas de menor prioridad a que tengan que desechar todo su trabajo con el recurso compartido.

Respecto de la sobrecarga del Sistema Operativo (4.2), sólo se analizó la situación con la garantía de que el recurso compartido estaba libre para poder ser bloqueado. El tiempo para el caso de techo de prioridades es mucho más alto que los demás, tal vez debido a la búsqueda inicial de la prioridad techo.

Como lineas futuras de investigación queda la opción de ver cuál es la sobrecarga cuando hay otras tareas en ejecución bloqueando el uso del recurso

[1] pthread_mutex_lock

compartido. También sería de interés analizar la sobrecarga que podría producir una implementación de restauración del recurso.

Referencias

1. RTwiki. http://rt.wiki.kernel.org/index.php/Main_Page.
2. L. Sha, R. Rajkumar, and J. P. Lehoczky. Priority inheritance protocols: An approach to real-time synchronization. *IEEE Trans. Comput.*, 39(9):1175–1185, September 1990.
3. Alan Burns and Andy Wellings. *Sistemas de tiempo real y lenguajes de programación*. Addison Wesley, Madrid [etc.], 2003.
4. Mars pathfinder. http://mars.jpl.nasa.gov/MPF/.
5. The risks digest volume 19: Issue 54. http://catless.ncl.ac.uk/Risks/19.54.html#subj6.
6. What really happened on mars? – authoritative account. http://research.microsoft.com/en-us/um/people/mbj/Mars_Pathfinder/Authoritative_Account.html.
7. Tarek Helmy and Syed S. Jafri. Avoidance of priority inversion in real time systems based on resource restoration. *IJCSA*, 3(1):40–50, 2006.
8. Victor Yodaiken. Against priority inheritance, 2002.
9. Priority inheritance in the kernel [LWN.net]. http://lwn.net/Articles/178253/.
10. How to use priority inheritance | embedded. http://embedded.com/design/configurable-systems/4024970/How-to-use-priority-inheritance.
11. J. Huang, J.A. Stankovic, K. Ramamritham, and D. Towsley. On using priority inheritance in real-time databases. In *Real-Time Systems Symposium, 1991. Proceedings., Twelfth*, pages 210–221, 1991.
12. Butler W. Lampson and David D. Redell. Experience with processes and monitors in mesa. *Commun. ACM*, 23(2):105–117, February 1980.
13. Andreu Carminati, Rômulo Silva de Oliveira, and Luís Fernando Friedrich. Implementation and evaluation of the synchronization protocol immediate priority ceiling in PREEMPT-RT linux. *Journal of Software*, 7(3), March 2012.
14. Priority inversion (windows). http://msdn.microsoft.com/en-us/library/windows/desktop/ms684831(v=vs.85).aspx.
15. FreeRTOS - market leading RTOS (real time operating system) for embedded systems supporting 34 microcontroller architectures. http://www.freertos.org/.
16. MaRTE OS home page. http://marte.unican.es/.
17. QNX operating systems, development tools, and professional services for connected embedded systems. http://www.qnx.com/.
18. RTAI - official website. https://www.rtai.org/.
19. RTLinux. http://es.wikipedia.org/w/index.php?title=RTLinux&oldid=64581733, March 2013. Page Version ID: 64581733.
20. Rtlinuxpro Victor Yodaiken and Victor Yodaiken. Temporal inventory and real-time synchronization in.
21. Wind river. http://www.windriver.com/.

Estudio sobre mediciones de Campos Electromagnéticos No Ionizantes

Jorge S. García Guibout[1/2]. Miguel Méndez Garabetti[1], Antonio Castro Lechtaler[3], Alfredo David Priori (estudiante becado)[1]
[1] Universidad del Aconcagua, [2] Instituto Tecnológico Universitario,
[3] Universidad Tecnológica Nacional
(jgarcia@itu.uncu.edu.ar, miguelmendezgarabetti@gmail.com, antonio.castrolechtaler@gmail.com, dalf_p@hotmail.com)

Abstract. En la actualidad muchas tecnologías de comunicaciones, medicina y del hogar se basan en la emisión de ondas electromagnéticas (TV, WiFi, telefonía celular, hornos a microondas, etc.). Éstas interactúan con nuestro cuerpo y no se tiene, aun, un acabado conocimiento de esa interacción. Esta situación ha provocado que la sociedad comience a preocuparse por las posibles consecuencias de estas emisiones, pidiendo mayores controles y reglamentaciones. Este trabajo busca conocer, en base al estado actual del conocimiento, la forma correcta de medir estos campos para que se puedan acompañar y justificar las decisiones sobre este tema que pueden llegar a ser necesario tomar en las zonas de influencia del Gran Mendoza y del país todo.

Keywords: radiación ionizante, radiación no ionizante, campo eléctrico, campo magnético, ondas electromagnéticas, mediciones.

1 Introducción

Los avances tecnológicos han producido cambios trascendentales y muchos de estos cambios son debidos a servicios que emplean ondas electromagnéticas, en especial los servicios inalámbricos. Esta situación lleva a que los seres vivos se vean expuestos constantemente e involuntariamente a los efectos de dichas radiaciones. Éstas pueden resultar perjudiciales para la salud (efectos negativos que pueden provocar riesgos en la salud).

Si bien se han realizado numerosas investigaciones para conocer los posibles efectos negativos que este tipo de radiaciones pueden causar en la salud en función a la intensidad con las que las mismas inciden en ellos, algunos de estos son conocidos[1], otros son aun controvertidos[2] por lo que estas investigaciones aún son insuficientes.

La ***Contaminación Electromagnética*** reconocida por la organización Mundial de la Salud, en primera instancia se enfocó en las antenas de televisión, antenas de radiodifusión, tanto AM y FM, líneas de alta tensión y otras fuentes de RNI [7]. El desplie-

[1] Como ser: calentamiento térmico, inducción de corriente eléctrica, etc.
[2]Como pueden ser: ciertos tipos de cáncer, alteraciones al sistema nervioso central, leucemia infantil, etc.

gue creciente de la telefonía móvil, según estimó la Unión Internacional de Telecomunicaciones - ITU [9] habría llegado en 2011 a 5.800 millones de suscriptores. Esta situación generó preocupación e incertidumbre, máxime teniendo en cuenta que este número tan importante de usuarios del servicio impacta directamente en la cantidad de antenas instaladas en ciudades y pueblos.

2 Radiación

La radiación podemos definirla como la propagación de energía, sea esta en forma de ondas electromagnéticas o partículas a través del espacio.

Toda radiación electromagnética está asociada con la emisión de fotones que son los responsables de la interacción electromagnética. Las características del fotón hacen que claramente podamos observar en la vida diaria fenómenos que son explicados tanto por su naturaleza corpuscular como ondulatoria.

El fotón es una partícula sin masa y sin carga eléctrica, lo que hace que viaje en el vacío a la velocidad de la luz. La ausencia de carga eléctrica hace que sea no-ionizante, es decir que por sí misma no puede alterar el equilibro de carga eléctrica por donde pase. Cada fotón se caracteriza por su energía que es directa y biunívocamente proporcional a su frecuencia, calculando la energía del fotón por la ecuación siguiente:

$$E = h\, f \text{ [Julios]} \qquad \textbf{(1)}$$

Donde: f es la frecuencia y h es la constante de Planck que es igual a $6{,}626 \times 10^{-34}$ [Js^{-1}].

La energía se suele medir en Julios, pero como la energía asociada a cada tipo de radiación electromagnética es débil, se emplea otra unidad de energía llamada eV (electronvoltio).

$$1eV = 1{,}602176462 \times 10^{-19}\ J \qquad \textbf{(2)}$$

La constante de Planck en dichas unidades será igual a $4{,}135567 \times 10^{-15}$ [eV].

En la naturaleza todas las partículas tienen un comportamiento dual: como partículas y como ondas. Así, algunos fenómenos de la radiación pueden ser entendidos como si fueran ondulaciones, y para otros fenómenos tendremos que concebirlos como si fueran partículas.

La realidad es que son ambas cosas, de allí el concepto de dualidad.

El comportamiento ondulatorio se manifiesta por medio de campos eléctricos y magnéticos perpendiculares entre sí que oscilan perpendicularmente a la dirección de propagación de la onda. Estas ondas electromagnéticas pueden ser descritas mediante sus parámetros característicos como lo son su frecuencia o longitud de onda y contenido de energía.

En función de la energía que tenga la onda electromagnética la radiación puede ser ionizante (RI) o no ionizante (RNI).

Una onda de radio, menor a 30 KHz hasta 1 GHz, poseen una energía desde $1{,}24x10^{-10}$ eV hasta $4{,}14x10^{-6}$ eV, la Luz Ultravioleta, 790 THz, con un energía de 3,3 eV y la Radiación Ionizante, 952 THz en adelante, tienen una energía mayor a 4 MeV.

3 Radiación Ionizante (RI) y NO Ionizante (RNI)

La radiación ionizante podemos definirla cómo: ***las radiaciones que por su frecuencia son capaces de entregar energía a los átomos de las sustancias como para romper los enlaces químicos, desprender un electrón y de esta manera crear un ion, e incluso interactuar con el núcleo del átomo*** [4]. Cuando un átomo pierde uno de sus electrones se dice que se ioniza, convirtiéndose en un ión o un catión y aún modificar la estructura del núcleo desprendiendo neutrones o protones.

Este es el caso de la radiación ultravioleta, los rayos x y los rayos gamma, siendo estos últimos los que pueden interactuar a nivel del núcleo. La radiación ionizante es producida por diversas fuentes como fuentes cósmicas externas (radiación cósmica), materiales radiactivos naturales contenidos en la corteza terrestre, en los ecosistemas y en el interior de los organismos vivos, los que pueden emitir, según sea el elemento, partículas Alfa y Beta, rayos Gamma y “radiación exótica” debida a materiales radioactivos producidos por el ser humano a partir de 1945 (fuentes bélicas y experimentales, fuentes civiles), aparatos que producen rayos X como energía residual, radiación solar cuya porción ultravioleta C no haya sido detenida por la alta capa de ozono (10^{16} a 10^{17} Hz).

Este tipo de radiación en su interacción con la materia puede causar daños en tejidos biológicos incluyendo efectos sobre el ADN (ácido desoxirribonucleico: material genético de los seres vivos), por tales motivos las aplicaciones que utilizan este tipo de radiación se utilizan en recintos aislados con importantes cuidados al medioambiente y del personal que opera la tecnología.

Por el contrario de la Radiación Ionizante, la Radiación No-Ionizante (RNI) podemos definirla como: ***las radiaciones que no poseen la suficiente energía, para desprender electrones de los átomos*** [4].

Este tipo de radiación se extiende desde las frecuencias muy bajas de la luz ultravioleta hasta las frecuencias extremadamente bajas como las del tendido eléctrico (ELF) y los campos magnéticos y eléctricos de naturaleza estática.

Su efecto principal es el incremento de la temperatura del material con el que interacciona. Esto es debido a que el fotón al interaccionar con la materia es como si ***chocara*** con ella, es decir, su energía pasa a la materia en forma de incremento de Energía Cinética.

La ionización se produce en forma abrupta a partir de un umbral de frecuencia y este umbral es una barrera de energía perfectamente definida, que es diferente en cada material.

Si bien este tipo de ondas electromagnéticas no pueden ionizar la materia incidida, si pueden causar otro tipo de efectos sobre la materia. De aquí es que podemos clasificar a los efectos de las RIN en:

- ✓ Efectos térmicos,
- ✓ Efectos no térmicos o biológicos.

El cuerpo humano posee mecanismos para regular de forma eficiente su temperatura, pero si la exposición a campos electromagnéticos es demasiado alta, el cuerpo podría no ser capaz de regular tal incremento, por este motivo es que los límites de exposición previenen un incremento de temperatura en el cuerpo humano de 1º C.

4 Tasa de Absorción Específica

Ya que los límites de exposición pueden ser establecidos en distintas unidades, para las frecuencias más bajas y hasta varios cientos de MHz, se suele utilizar la intensidad del campo eléctrico expresada en V/m, la intensidad de campo magnético en A/m o la densidad de potencia expresada mW/cm^2 o W/m^2.

Pero existe un parámetro dosimétrico ampliamente utilizado el cual se denomina "tasa de absorción específica" o ***SAR Specific Absorption Rate***, el cual se define como: ***La derivada del aumento de la energía, ∂W, absorbida o disipada en un elemento de masa ∂m, contenida en un elemento de volumen ∂V, cuya densidad es ρ***. Puede ser expresado analíticamente como [1]:

$$SAR = \frac{\partial}{\partial t}\frac{\partial W}{\partial m} = \frac{\partial}{\partial t}\frac{\partial W}{\partial (\partial V)}\left[\frac{mW}{g}\right] \qquad (3)$$

En la ecuación siguiente se puede observar que el SAR es directamente proporcional al aumento local de la temperatura:

$$\frac{\partial T}{\partial t} = \frac{SAR}{C_p}\left[{}^{\circ}C/_{s}\right] \qquad (4)$$

donde T es la temperatura en grados Celsius, y Cp es el calor específico del tejido (J/kg °C).

O sea, que la tasa de absorción específica es la medida de la cantidad de energía de radiofrecuencia que es absorbida por los tejidos en el cuerpo humano y se expresa en W/kg.

5 Efectos en la Salud de las Radiaciones No Ionizantes

La preocupación por este nuevo tipo de contaminación se ha acentuado con la aparición de la telefonía móvil, la instalación y permanente funcionamiento de una gran cantidad antenas fijas que operan en el rango de las microondas, y la multiplicación de miles de pequeñas antenas móviles que emiten y reciben estas señales (teléfonos celulares, terminales del sistema, o teléfonos móviles, etc.) [4]. Es por ello que en

diversos ámbitos han continuado las investigaciones respecto a los efectos de ellas tanto en el ambiente como en el cuerpo humano.

De acuerdo a estudios realizados, la energía radiante puede ser absorbida por el cuerpo humano mediante tres procesos diferentes: ***efecto antena, absorción de señal*** (proceso de absorción relacionado con la constante dieléctrica y el tipo de conductividad del tejido, los cuales son diferentes a distintos valores de frecuencia), ***absorción biofísica*** (involucra la absorción resonante por sistemas biológicos como el cerebro o las células).

Algunos de los bioefectos y sus mecanismos dependen del rango del espectro de frecuencias de las ondas.

Ésta se puede dividir en frecuencia extremadamente baja, que es toda frecuencia menor a 30 KHz, y radiofrecuencia y microondas, que incluye frecuencias hasta 300 GHz, que actúan a nivel de las estructuras de las células que están conformadas por moléculas y átomos cargados que pueden cambiar su orientación y movimiento cuando se encuentran expuestos a una fuerza electromagnética

Se ha visto que la radiación de los campos eléctrico y magnético de las frecuencias extremadamente bajas pueden existir separadamente el uno del otro. Normalmente, la discusión sobre los efectos se restringe normalmente al campo magnético que es producido por corrientes alternas o campos variantes en el tiempo, cuya intensidad y dirección cambien de forma regular, ya que los campos eléctricos son fácilmente apantallados [13].

Por otra parte, la exposición a radiaciones de frecuencias extremadamente bajas ocurre a distancias mucho menores que la longitud de onda. Esto tiene importantes implicancias porque bajo tales condiciones se tratan como componentes independientes.

La situación es sustancialmente diferente de la que ocurre en la radiación de campos de alta frecuencias, en donde los campos eléctrico y magnético están indisolublemente unidos. Esta es la razón por la que las investigaciones se han centrado en los efectos de un campo o el otro, en frecuencias extremadamente bajas.

La interacción del campo electromagnético con sistemas vivos que se ha propuesto teóricamente es la habilidad del campo magnético para estimular corrientes en las membranas de las células y en los fluidos de los tejidos, que circulan en un lazo cerrado que descansa en un plano perpendicular a la dirección del campo magnético. Por tanto, en el interior de un medio biológico se inducen corrientes y campos eléctricos debido al campo magnético.

Una posible interacción bajo investigación ***es que la exposición a campos de frecuencias extremadamente bajas suprime la producción de melatonina***, que es una hormona producida por la glándula pineal localizada en una zona profunda del cerebro.

La melatonina se produce principalmente por la noche y se libera al cuerpo a través del flujo sanguíneo. Ella llega a casi todas las células del cuerpo humano, estimulando

el sistema inmune, preserva el ADN, las proteínas y los lípidos de daños oxidativos al neutralizar los radicales libre que pueden causar daños estructurales [10].

Además regulan otras actividades como los ciclos menstruales femeninos, el ritmo cardíaco, el sueño, el estado de ánimo y la genética y es esencial para el sistema inmunológico, protegiendo al cuerpo de infecciones y de las células cancerosas.

Diversos estudios han encontrado reducción de melatonina en células animales y personas expuestos a campos de frecuencias extremadamente bajas siendo un efecto que depende fuertemente del período de exposición y de la intensidad del campo

En cuanto a los efectos biológicos de la RF y microondas se han desarrollado un número significativo de estudios. Estos exploran la posible relación entre la exposición a la radiación de campos RF y microondas y las enfermedades, incluyendo el cáncer; pero todavía deberá pasar un tiempo hasta que se tengan los resultados finales de la mayoría de los estudios.

Básicamente, existen dos tipos de efectos biológicos a estas frecuencias

- ✓ ***Efectos térmicos***: ocurren cuando la radiación en cuestión posee suficiente energía como para ocasionar un incremento de temperatura medible.
- ✓ ***Efectos no térmicos***: es una línea de investigación en pleno desarrollo. Podemos decir que se registran efectos biológicos a niveles SAR muy por debajo de los 0,08 W/kg y a densidades de potencia minúsculas de 0,0004μW/cm^2.

Es muy importante remarcar que los estándares tanto del ***ICNIR - International Commission on Non-Ionizing Radiation Protection***, la ***Organización Mundial de la Salud*** y la ***Unión Europea*** se basan, en su mayoría, en efectos térmicos de naturaleza irreversible para exposiciones a corto plazo.

6 Mediciones de las ondas Electromagnéticas

La investigación en el área de los bioefectos de la radiación electromagnética ionizante[3] fue anterior a su utilización, lo que permitió reducir los riesgos y por lo tanto aumentar la utilización de dispositivos nucleares generadores de energía, así como también de aquellos derivados de la tecnología de radiación X (medicina, etc.)

Los efectos de las radiaciones electromagnéticas no ionizantes de radiofrecuencias son motivo de preocupación, por lo tanto el problema de la dosimetría es muchísimo más complicado en este otro caso[4], que en el de la radiación electromagnética ionizante.

Es obvio que los estándares de protección contra la radiación de radiofrecuencias deben expresarse en términos de la intensidad del campo Eléctrico (E), campo Magnético (H) y Densidad de Potencia en el espacio libre (S, como se lo llama en la resolución 3690 de la CNC [16]).

[3] Rayos X y gamma.
[4]Radiofrecuencias.

El propósito de la medición o prospección de radiación es medir los campos E, H, y S en el ambiente donde el hombre puede estar eventualmente expuesto y comparar esas mediciones con los estándares de niveles permisibles de exposición establecidos

Con el fin de poder determinar las condiciones de exposición mínima a las que son expuestas la población y aquellos agentes que trabajan directamente sobre equipamiento emisor de señales electromagnéticas, el Ministerio de Salud y Acción Social de la Nación estableció por medio de la resolución 202/95 [17] los niveles mínimos a los que pueden ser expuestos la población en general y a personas expuestas por su ocupación.

En base a esta resolución del Ministerio de Salud y Acción Social de la Nación la Comisión Nacional de Comunicaciones (CNC) emite la Resolución 3690/2004 [16], que abarca a las 269/2002 y la 217 CNC/2003 de la misma repartición, donde se establecen las formas en que deben medirse los niveles de campos eléctricos, magnéticos y/o densidades de potencia en el áreas bajo interés para corroborar lo dictaminado por el Ministerio de Salud y Acción Social de la Nación..

Se presentan en la tabla 1 los valores propuestos por la Resolución 202/95 [17].

Tabla 1: Niveles propuestos por la Resolución 202/95 de MSAS

Rango de frecuencia f (MHz)	**Densidad de Potencia equivalente de onda plana S (mW/cm^2)**	**Campo Eléctrico E (V/m)**	**Campo Magnético H (A/m)**
0,3 - 1	20	275	0,73
1 - 10	20/f2	275/f	0,73/f
10 - 400	0,2	27,5	0,073
400 - 2000	f/2000	1,375 f/2	-
2000 - 100.000	1	61,4	-

Para la toma de las mediciones, la resolución de la CNC define un campo cercano y un campo lejano, existentes en las proximidades de una antena.

En la ***región del campo lejano***, a una distancia mayor que 3λ de la antena, el campo predominante es del tipo onda plana, es decir una distribución localmente uniforme de la intensidad de campo eléctrico y de la intensidad de campo magnético en planos transversales a la dirección de propagación.

La ***región de campo cercano*** se subdivide a su vez en la región de campo cercano reactivo, que es más próxima al elemento radiante y la región de campo cercano ra-

diante, en la que el campo de radiación predomina sobre el campo reactivo, pero que no es sustancialmente del tipo onda plana y tiene una estructura compleja.

El campo Eléctrico se expresará en V/m y el campo Magnético en A/m, la Densidad de Potencia (S) se expresa en mW/cm^2.

En una onda plana estos parámetros están relacionados por medio de la impedancia del espacio libre ($Z_0 = 377\ \Omega$), por lo tanto con la medición de algunos de los campos será suficiente para obtener el resto por medio de la ecuación:

$$S = \frac{E^2}{Z_0} = H^2 Z_0 \tag{5}$$

En el caso de mediciones en el campos cercanos, las componentes de los campos eléctricos (E) y magnéticos (H) son generalmente desconocidas. Por ello, se deberá, en todos los casos, realizar la medición de dichos campos en forma separada.

Es necesario introducir las definiciones, que forman parte de de la resolución de la CNC, de ***Emisión*** que es la radiación producida por una única fuente de radiofrecuencia, y la de ***Inmisión*** que es la radiación resultante del aporte de todas las fuentes de radiofrecuencias cuyos campos están presentes en el lugar.

También definimos ***Potencia Radiada Aparente - PRA*** como el producto de la potencia suministrada a la antena por la ganancia de antena, en una dada dirección, relativa a un dipolo de media onda y ***Potencia Isotrópica Radiada Equivalente-PIRE como*** el producto de la potencia suministrada a una antena por la ganancia de antena, en una dada dirección, relativa a una antena isotrópica.

El procedimiento de evaluación para aquellas estaciones cuyas características de radiación impliquen la consideración del campo lejano, la evaluación de los valores de radiaciones no ionizantes (RNI) para el caso de una antena única, las predicciones de densidad de potencia se pueden realizar a partir de las siguientes ecuaciones, que si bien son solamente válidas para los cálculos en el campo lejano de una antena, pueden utilizarse para predecir el peor de los casos,

$$S = \frac{PRA * 1{,}64 * 2{,}56 * F^2}{4 * \pi * r^2} \tag{6}$$

Donde:

PRA se considera en W

F es la atenuación en veces de la radiación para un cierto ángulo de incidencia en el plano vertical, que si es desconocido toma un valor igual a 1

2.56 es un factor de reflexión empírico, que tiene en cuenta que se puedan adicionar campos reflejados en fase.

r es la distancia desde la antena

O por medio de

$$S = \frac{PIRE * 2{,}56 * F^2}{4 * \pi * r^2} \tag{7}$$

Donde: PIRE se considera en W

De estas ecuaciones se puede despejar la distancia mínima a la antena, r, con los valores de densidad de potencia establecidos en la tabla 1 sobre límite de exposición poblacional.

En el caso que se determine que se superan los valores límites establecidos en la tabla 1, se deberán llevar acabo mediciones según el protocolo que se detalla a continuación.

En base a las características del sistema irradiante y la de los emisores se determinaran los puntos de mayor riesgo tanto externos al predio de la antena como internos al mismo. Se deberá tener en cuenta la topología, edificaciones y superficies reflectoras del lugar.

La medición se efectuará en los puntos de mayor acceso por parte del público. En sistemas omnidireccionales se deberán seleccionar 16 puntos como mínimo y para sistemas direccionales se deberán adoptar un mínimo de 4 puntos sobre la dirección de máxima propagación, los 12 puntos restantes deberán ubicarse en función de las características del lóbulo de radiación de dicha fuente. Todos estos puntos serán función de la longitud de onda del sistema emisor.

El tipo de instrumento establecido en la resolución son equipos de banda ancha que responden uniforme e instantáneamente a un amplio rango de frecuencias y no son sintonizables. Éstos se emplean con sondas de medición de E y H del tipo isotrópico, para una respuesta independiente de la orientación de la sonda. Los mismos son utilizados para la medición de inmisión.

También se menciona instrumentos de banda angosta que operan sobre un amplio rango de frecuencias, pero su ancho de banda instantáneo de medición se reduce a anchos de banda estrechos. Este tipo de dispositivos debe sintonizarse a la frecuencia de interés y utilizarse con antenas aptas para los distintos rangos de frecuencia de medición. Son utilizados para la medición de emisión y proporcionan información de la frecuencia bajo análisis.

La secuencia de medición establecida indica que en primer término se medirá inmisión. Si los valores obtenidos superaren los máximos permisibles más estrictos dados en la tabla 1, se continuará midiendo la emisión de cada estación.

La medición de inmisión tiene por objeto obtener el nivel pico máximo de los campos eléctrico, magnético o de la densidad de potencia, a lo largo de una línea vertical que represente la altura del cuerpo humano en el punto de medición.

Estas mediciones comienzan a 20 cm por encima del suelo hasta una altura de 2 m a una velocidad constante. Si el valor pico máximo de dichas mediciones resulta inferior al 50% de la ***Máxima Exposición Permitida -MEP*** más estricta, se registrará como valor de ese punto. Si dicho valor supera el citado 50% de la MEP más estricta, se deberá realizar una medición con promediado temporal de 6 minutos.

En caso que los resultados obtenidos en las mediciones de inmisión superen los límites de la tabla 1, se deberá proceder a la medición de emisión a fin de evaluar los aportes individuales de cada una de las fuentes emisoras de radiaciones no ionizantes.

Se medirá la intensidad de campo producida por la estación a verificar sobre cada uno de los puntos de medición seleccionados, por medio de instrumentos de banda angosta asociados con antenas de polarización lineal. A tal efecto podrán utilizarse dos métodos alternativos:

a) Orientar la antena en tres direcciones ortogonales entre sí (x, y, z) obteniéndose las componentes de campo respectivas. Los valores cuadráticos de intensidad de campo eléctrico y/o magnético se obtendrán de la suma de los cuadrados de las correspondientes componentes de campo ortogonales, como se observa en las siguientes ecuaciones:

$$E^2 = E_X^2 + E_Y^2 + E_Z^2 \qquad (8)$$

$$H^2 = H_X^2 + H_Y^2 + H_Z^2$$

b) Orientar la antena en la dirección de máxima señal. Este método es también aplicable a una antena de apertura

7 Conclusiones

La permanente y rápida evolución de nuevas tecnologías que utilizan campos electromagnéticos para brindar servicios cada vez más útiles y novedosos, no ha permitido que en forma simultánea ***se hayan realizado acabadamente las investigaciones de los posibles efectos negativos sobre personas y ecosistemas antes de su masificación.***

De acuerdo con la bibliografía analizada, se está trabajando fuertemente en la investigación en torno a este fenómeno, tanto en nuestro país como así también en el resto del mundo.

Es importante ***poder medir las magnitudes de los campos magnéticos y eléctricos a los que nos vemos expuestos***, de manera que la sociedad esté segura que se cumplen las normativas de seguridad de exposición y tener las herramientas para solicitar las correcciones que sean necesarias.

Para ello es necesario conocer las especificaciones nacionales y compararlas con regulaciones internacionales para mejorar en la toma de estas mediciones, tratando de replicarlas en el laboratorio, para después poder adaptarlas a nuestras topología, dis-

tribución poblacional, particularidad de nuestra estructura edilicia, entre otras características propias de nuestra región.

En este sentido es importante destacar el aporte del Laboratorio de Investigación Aplicada y Desarrollo (LIADE) de la Universidad Nacional de Córdoba en el estadio y difusión de este tema [11,12].

8 Bibliografía

1. International Commission on Non-Ionizing Radiation Protection: Recomendaciones para limitar la exposición a campos eléctricos, magnéticos y electromagnéticos (hasta 300 GHz): http://www.icnirp.de/documents/emfgdlesp.pdf.
2. Organización Mundial de la Salud: Campos electromagnéticos y salud pública: radares y salud humana:http://www.who.int/pehemf/publications/facts/fs226/es/
3. Mobile Phone Simulations with Human Head and Hand Models; Computer Simulation Technology; 2011 CST AG, http://www.cst.com/Content/Applications/ Article/Mobile+Phone+Simulations+with+Human+Head+and+Hand+Models,
4. Capparelli M., Mata N., Montenegro R., Aliciardi M.(2008). El ambientalismo II, la Electropolución, Contaminación por antenas de telefonía celular., editorial Ediciones del País.
5. Agencia Internacional sobre Investigación del Cáncer: http://www.iarc.fr/ index.php
6. Organización Mundial de la Salud; http://www.who.int/about/es/
7. IARC classifies radiofrequency electromagnetic fields as possibly carcinogenic to humans, Press Release N° 208 , 31 de mayo 2011. http://www.iarc.fr/en/mediacentre/pr/2011/pdfs/pr208_E.pdf
8. IEEE Std C95.3-2002 (Revision of IEEE Std C95.3-1991). IEEE Recommended Practice for Measurements and Computations of Radio Frequency Electromagnetic Fields With Respect to Human Exposure to Such Fields, 100 kHz–300 GHz. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8351
9. ITU: http://www.itu.int/es/about/Pages/default.aspx
10. Instituto Latinoamericano de la Comunicación Educativa: http://bibliotecadigital.ilce.edu.mx/sites/ciencia/volumen2/ciencia3/107/htm/paraatra.htm
11. Bruni R., Vanilla O., Taborda R., Evaluación de radiación electromagnética de antenas, L.IADE (Laboatorio de Investigación Aplicada y Desarrollo) – Universidad Nacional de Córdoba, www.liade.efn.uncor.edu/
12. Bruni R., Vanilla O., Taborda R, Evaluación de radiación electromagnética de fuentes no naturales, L.IADE (Laboatorio de Investigación Aplicada y Desarrollo) – Universidad Nacional de Córdoba, www.liade.efn.uncor.edu/
13. Cátedra Radiaciones, Universidad Nacional de Entre Ríos. http://www.bioingenieria.edu.ar/academica/catedras/radiaciones/Descargas/Unidad 1.pdf
14. Portela A., Skvarca J., Matute Bravo E., Loureiro A., Prospección de la radiación electromagnética no Ionizante, Vol.I: Manual de estándares de seguridad para la exposición a radiofrecuencias comprendidas entre 100 KHz y 300 GHz, Dirección

Nacional de Calidad Ambiental Secretaria de Salud Ministerio de Salud y Acción Social.
15. Portela A., Skvarca J., Matute Bravo E., Loureiro A.,Prospección de la radiación electromagnética no Ionizante, Vol.II: Radiación de radiofrecuencias: Consideraciones biofísicas, biomédicas y criterios para el establecimiento de estándares de exposición.
16. Comisión Nacional de Comunicaciones (CNC) (2004). Resolución 3690/2004 (Boletín oficial Nº 30.524, 10/11/04).
17. Ministerio de Salud y acción Social de la Nación (1995). Resolución 202/95 (EXP Nº2002-17655- 94-04).

Selección sub-óptima del espectro asociado a la matriz de afinidad

Luciano Lorenti, Lucía Violini, Javier Giacomantone

Instituto de Investigación en Informática (III-LIDI),
Facultad de Informática - Universidad Nacional de La Plata - Argentina.
La Plata, Buenos Aires, Argentina.
{llorenti,lviolini,jog}@lidi.info.unlp.edu.ar

Resumen. En este artículo se presenta un método de agrupamiento espectral que incorpora un etapa de selección sub-óptima de características. Los métodos de agrupamiento espectral tienden a determinar la estructura subyacente en un conjunto de patrones, donde otros métodos convencionales por la disposición y características particulares de los agrupamiento, no obtienen los resultados esperados. En este trabajo se propone utilizar un método particular de selección de características que tiene como objetivo determinar el mejor conjunto de autovectores de la matriz de afinidad normalizada. La determinación correcta del subconjunto de autovectores más relevantes, puede ser utilizada para mejorar las características de las particiones generadas. El método es evaluado con datos sintéticos simulando estructuras de datos específicas, y en datos reales obtenidos con una cámara de tiempo de vuelo.

Palabras clave: Clustering espectral, Selección de características, Visión por Computadora, Robótica.

1 Introducción

Una etapa general de un sistema de reconocimiento automático de patrones es la de reducción de dimensión [1][2]. La reducción del número de características puede realizarse en modo supervisado o no supervisado y las técnicas empleadas pueden ser clasificadas como de selección o de extracción de características [3]. En selección de características se definen criterios para elegir el mejor subconjunto de características de un conjunto original [4][5], y en extracción de características se generan nuevas características mediante transformaciones lineales o no lineales del mismo. El problema particular para el cual el sistema es diseñado y el tipo de datos que define los objetos o fenómenos tratados, determinará si se incluirán las dos etapas, una de extracción y una de selección, o solamente una etapa. El diseño de un módulo de selección de características involucra múltiples consideraciones sobre el conjunto de patrones disponibles, como valores atípicos, imputación de valores, tipo de normalización y compromisos en el número óptimo de características, dadas por el fenómeno de pico [6]. Sistemas con restricciones

temporales o donde el número de características es mucho mayor que el número de muestras son alguno de los casos en que se plantea utilizar sistemas de reducción de dimensión. En este trabajo se propone un método en el cual se seleccionan autovectores en el contexto de un método de agrupamiento (clustering) y no como una etapa clásica en el diseño de un sistema de clasificación. Cuando la estructura de los datos no corresponde a regiones convexas, no es lineal o cuando los métodos clásicos de agrupamiento, jerárquicos o particionales, no obtienen resultados satisfactorios, una alternativa son los métodos de agrupamiento espectral (spectral clustering) [7][8][9]. Las técnicas de agrupamiento espectral utilizan los autovectores de la matriz de afinidad, o de matrices derivadas, para generar una partición del conjunto de muestras en agrupamientos disjuntos, que presenten valores altos de la medida de semejanza adoptada para patrones en un mismo conjunto, y bajos para patrones de conjuntos diferentes. En general el valor de los correspondientes autovalores determina un criterio que permite establecer la prioridad de los autovectores, esta no necesariamente genera la mejor partición del espacio muestral. Es posible entonces aplicar técnicas de selección de características para determinar qué combinación de autovectores genera la mejor partición [10][11][12]. En este artículo se propone un método que combina el potencial de los métodos de agrupamiento espectral con un método específico de selección sub-óptima de características [13]. Se presentan resultados experimentales utilizando datos artificiales e imágenes de rango reales. En particular se utiliza una cámara de tiempo de vuelo [14], utilizada en aplicaciones generales de visión por computadora [15] y robótica [16][17].

En la sección 2 se describe el método de selección sub-óptima de búsqueda secuencial flotante hacia adelante adoptado. En la sección 3 se describe método de cortes normalizados y en la sección 4 se presenta el método propuesto. En la sección 5 se muestran resultados experimentales. Finalmente, en la sección 6 se presentan las conclusiones.

2 Selección Sub-óptima de características

2.1 Selección de Características

El problema de selección de características consiste en, dado un conjunto Y de D características determinar cuál es el subconjunto X de tamaño $d < D$ que genera la mayor contribución a la discriminación entre clases. Se puede plantear el problema en términos de la optimización de una función criterio J para un subconjunto de tamaño d.

$$J(X) = \max_{Z \subset Y, |Z|=d} J(Z)$$

El método de selección determinará una función criterio y el algoritmo de búsqueda adecuado para el problema analizado. Una vez determinada la función criterio es necesario definir un algoritmo de búsqueda que tenga como objetivo un compromiso entre optimización y complejidad que resulte viable. El método sub-óptimo

adoptado en este artículo es el método de búsqueda secuencial flotante hacia adelante (SFFS - Sequential Forward Floating Selection) [13][18] y la función criterio adoptada basada en matrices dispersas. El método SFFS fue propuesto para evitar el problema de anidamiento que genera el método de búsqueda secuencial hacia adelante (SFS - Sequential Forward Selection) [19].

2.2 Búsqueda Secuencial Flotante hacia Adelante

Dado el conjunto completo de D mediciones $Y = \{y_j \mid j = 1, \ldots, D\}$, seleccionar k características, para formar el conjunto $X_k = \{x_j \mid j = 1, \ldots, k, \quad x_j \in Y\}$, que optimice la función criterio $J(X_k)$ correspondiente.

Inicialización:
$X_0 := \emptyset; \quad k := 0$
en la práctica se puede comenzar con k = 2 por aplicar SFS dos veces.

Paso 1 *(Inclusión):*
$x^+ := \arg \max_{x \in Y - X_k} J(X_k + x)$
x^+ es la característica más significativa con respecto a X_k.

$X_{k+1} := X_k + x^+; \quad k := k + 1$

Paso 2 *(Exclusión Condicional):*
$x^- := \arg \max_{x \in X_k} J(X_k - x)$
x^- es la característica menos significativa en X_k.

`if` $J(X_k - \{x^-\}) > J(X_{k-1})$ `then`
  $X_{k-1} := X_k - x^-; \quad k := k - 1$
  `ir a Paso 2`
`else`
  `ir a Paso 1`

Terminación:
Parar cuando k es igual al número de características requerido

2.3 Medida de Semejanza entre clases

Debido al uso recursivo de la función criterio, se propone utilizar la siguiente medida de separación entre clases basada en matrices dispersas.

$$J = tr\{S_w^{-1} S_m\} \quad \text{donde} \quad S_w = \sum_{i=1}^{M} P_i \Sigma_i \quad \text{y} \quad S_m = E[(x - \mu_0)(x - \mu_0)^T]$$

siendo M el número de clases y P_i la probabilidad a priori de cada clase.

3 Agrupamiento Espectral

El algoritmo de cortes normalizados propuesto por Shi y Malik [20] modela la segmentación de imágenes como un problema de partición de un grafo. Un grafo G=(V,E) está formado por un conjunto de vértices V y un conjunto de aristas E que relacionan elementos de V. Con el objetivo de construir grafos a partir de imágenes, los vértices son generados a partir de los pixeles que la constituyen. El conjunto de las aristas E está formado por elementos que denotan la semejanza entre los pixeles. Dado un grafo pesado no dirigido $G = (V, E)$, donde V son los nodos y E son las aristas. Sea A,B una partición de un grafo: $A \cup B = V, A \cap B = \emptyset$. La semejanza entre dos grupos es llamada *cut*

$$cut(A, B) = \sum_{i \in A, j \in B} w(i, j)$$

donde $w(i, j)$ es el peso de la arista que conecta el vértice i con el vértice j. El criterio propuesto por Shi y Malik utiliza un criterio de semejanza normalizado para evaluar la partición:

$$Ncut(A, B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

Una de las ventajas más importantes para usar el criterio de cortes normalizados es que se puede obtener una buena aproximación de la partición óptima de forma muy eficiente. Sea $W_{ij} = w(v_i, v_j)$ la matriz de pesos del grafo y sea D la matriz diagonal de forma que $D_{ii} = \text{grado}(v_i) = \sum_{v_j \in V} w(v_i, v_j)$
Shi y Malik demostraron que una partición óptima se puede obtener calculando:

$$y = \arg\min Ncut = \arg\min_{y} \frac{y^T(D - W)y}{y^T Dy}$$

donde y es un vector indicador binario que especifica a qué grupo pertenece cada pixel. Relajando el carácter discreto de y, la ecuación anterior puede ser aproximada resolviendo el sistema de autovalores generalizado:

$$(D - W)y = \lambda Dy$$

El segundo autovector de este sistema es la solución real del problema discreto de cortes normalizados.

3.1 Ncut simultáneo de k-vías

Se define el corte normalizado simultáneo de k-vías que da como resultado k segmentos en una sola iteración de la siguiente forma:

$$Ncut(A_1, A_2, \ldots, A_k) = \frac{cut(A_1,A_1)}{assoc(A_1,V)} + \frac{cut(A_2,A_2)}{assoc(A_2,V)} + \ldots + \frac{cut(A_n,A_n)}{assoc(A_n,V)}$$

Sea $L = (D - V)$, dado un vector indicador v de un sub-grafo A_j tal que

$$v_i = \begin{cases} \frac{1}{\sqrt{assoc(A_n, V)}} & \text{si } i \in A_j \\ 0 & \text{si } i \notin A_j \end{cases}$$

resulta que

$$v_i^T L v_i = \frac{cut(A_j, A_j)}{assoc(A_j, V)}$$

Sea H la matriz formada por k vectores indicadores puestos en columnas, minimizar $Ncut(A_1, A_2, \ldots, A_k)$ es equivalente a minimizar:

$$\min_{A_1, A_2, \ldots A_k} \text{Tr}(H^T L H) \text{ sujeto a } H^T D H = I$$

Se puede hallar eficientemente una aproximación continua de los autovectores discretos que minimizan la traza. Estos son los autovectores correspondientes a los k autovalores más pequeños del sistema de autovalores $(D - W)u = \lambda D u$.

4 Método Propuesto

Sea $M \in R^{n \times m}$ una matriz que contiene n patrones de m características.

1. Se construye la matriz de afinidad W de forma que

   $$W(i, j) = w(v_i, v_j)$$

   Para imágenes se utiliza:

   $$w(v_i, v_j) = e^{\frac{-\|F(i) - F(j)\|_2}{\alpha_I}} * \begin{cases} e^{\frac{-\|X(i) - X(j)\|_2}{\alpha_X}} & \text{si } \|X(i) - X(j)\|_2 < r \\ 0 & \text{c. c.} \end{cases}$$

   donde $X(i)$ es la locación espacial (x, y) del pixel i y $F(i)$ es el nivel de intensidad del pixel i.

2. Se calcula la matriz diagonal D tal que $D(i, i) = \text{grado}(v_i) = \sum_{v_j \in V} w(v_i, v_j)$
3. Sea H la matriz formada por los k autovectores correspondientes a los autovalores más pequeños del sistema de autovalores $(D - W)u = \lambda D u$.
4. Considerando las columnas de H como características y las filas como patrones se aplica el método de selección sub-óptima descripto en la sección 2.
5. Se aplica el algoritmo k-medias sobre el conjunto de patrones utilizando las mejores características.

5 Resultados Experimentales

En esta sección se presentan resultados experimentales preliminares del método propuesto aplicado a imágenes simuladas y a imágenes reales. Las evaluaciones comparativas del método han sido realizadas con distintos conjuntos de imágenes de rango, y con diferentes parámetros en la adquisición de las mismas. Las capturas reales fueron obtenidas utilizando la cámara de tiempo de vuelo MESA SwissRanger SR4000 [14] y las imágenes simuladas utilizando el software Blensor [21].

(a) Imagen de rango (b) Clustering espectral (c) Método propuesto

Fig. 1. Resultados sobre imagen de rango simulada

En la figura 1(a) se muestra una imagen de rango simulada compuesta por cuatro objetos equidistantes del plano focal. La figura 1(b) presenta la segmentación obtenida utilizando las tres primeras columnas de H. En la figura 1(c) se muestra la segmentación obtenida utilizando las columnas de H especificadas en la tabla 1 seleccionadas por el método propuesto.

(a) Imagen de rango (b) Clustering espectral (c) Método propuesto

Fig. 2. Resultados sobre imagen de rango real

La figura 2 muestra imágenes del experimento realizado sobre una imagen real obtenida utilizando la cámara de tiempo de vuelo MESA SR4000. En la figura

2(a) se muestra una imagen de rango correspondiente a dos objetos. La figura 2(b) presenta la segmentación obtenida utilizando clustering espectral. En la figura 2(c) se muestra la segmentación obtenida utilizando el método propuesto. En la tabla 1 se presentan los parámetros utilizados y las columnas de H seleccionadas.

(a) Imagen de rango (b) Clustering espectral (c) Método propuesto

Fig. 3. Resultados sobre imagen de rango real

La figura 3 muestra imágenes del experimento realizado sobre otra imagen real. En la figura 3(a) se muestra una imagen de rango correspondiente a dos objetos con niveles de profundidad similares. La figura 3(b) presenta la segmentación obtenida utilizando clustering espectral. En la figura 3(c) se muestra la segmentación obtenida utilizando el método propuesto. En la tabla 1 se presentan los parámetros utilizados y las columnas de H seleccionadas.

	Método	Columnas de H	r	α_I	α_X
Prueba 1	Espectral	1,2,3	4	8	2
	Propuesto	1,2,7	4	8	2
Prueba 2	Espectral	1,2,3	4	0.02	4
	Propuesto	5,7,10	4	0.02	4
Prueba 3	Espectral	1,2	4	0.045	4
	Propuesto	4,5	4	0.045	4

Tabla 1.

6 Conclusiones

En este artículo se propone utilizar un método de selección de características sub-óptimo embebido en un método de clustering espectral. Los resultados experimentales preliminares sobre datos simulados y reales permiten concluir que la selección adecuada del conjunto de autovectores de la matriz de afinidad mejora la determinación de la estructura subyacente en los patrones de entrada ya sea en datos simulados como reales. El método propuesto muestra potencial para su aplicación en imágenes de rango y de intensidad para problemas que presenten restricciones temporales, si se utilizan cámaras con características similares a la SR4000 de bajo tiempo de adquisición. El método es sensible al ajuste de los principales parámetros del mismo, por lo tanto como parte del trabajo futuro es necesario analizar el comportamiento de los mismos con respecto a los agrupamientos generados. Un segundo aspecto involucra evaluar el mecanismo de selección en otros métodos espectrales y realizar una evaluación comparativa de los resultados para estructuras de agrupamientos complejas.

Referencias

1. M .A. Carreira-Perpinan. A review of dimension reduction techniques. Technical report CS-96-09, Department of Computer Science, University of Sheffield, 1997.
2. I. K. Fodor. A survey of dimension reduction techniques. Technical report URL-ID-148494, Center for Applied Scientific Computing, Lowrence Livemore Laboratory, 2002.
3. P. Narendra, K. Fukunaga. A branch and bound algorithm for feature subset selection. IEEE Transactions on Computers, vol. C-26, no. 9, pp.917.922, 1977.
4. A. Blum, P. Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, vol. 97(1-2), pp. 245-271, 1997.
5. M. Kudo, J. Sklansky. Comparison of algorithms that select features for pattern classifiers. Pattern Recognition, vol. 33, pp. 25-41, 2000.
6. R. E. Bellman. Adaptive Control Processes: A Guided Tour. Princeton University Press, 1961.
7. J. Shi, J. Malik, Normalized cuts and image segmentation. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 731-737. 1997
8. P. Perona, W. T. Freeman. A factorization approach to grouping. ECCV, pp. 655-670. 1999.
9. A. Y. Ng, M. I. Jordan, Y. Weiss. On Spectral Clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, vol. 14, 2002.
10. T. Xiang, S. Gong. Spectral clustering with eigenvector selection. Pattern Recognition Vol. 41, 1012-1029, 2008.
11. F. Zhao, L. Jiao, H. Liu, X. Gao, M. Gong. Spectral clustering with eigenvector selection based on entropy ranking. Neurocomputing Vol. 73, pp. 1704-1717, 2010.
12. S. A. Toussi, H. S. Yazdi. Feature Selection in Spectral Clustering. International Journal of Signal Processing, Image Processing and Pattern Recognition vol. 4, No. 3, pp. 179-194, 2011.
13. P. Pudil, J. Novovicova, J. Kittler. Floating search methods in feature selection. Pattern Recognition Letters vol. 15, pp. 1119-1125, 1994.

14. M. Cazorla, D. Viejo, C.Pomares. Study of the SR 4000 camera. XI Workshop de Agentes Físicos, Valencia, 2010.
15. A. A. Dorrington, C. D. Kelly, S. H. McClure, A. D. Payne, M. J. Cree. Advantages of 3D Time of Flight Range Imaging Cameras in Machine Vision Applications. 16^{th} Electronics New Zealand. Dunedin, New Zealand, 95-99, 2009.
16. S. May, B. Werner, H. Surmann, K. Pervolz. 3D time-of-flight cameras for mobile robotics. IEEE International Conference on Intelligent Robots and Systems, pp.790-795, 2006.
17. A. Prusak, O. Melnychuk, H. Roth, I. Schiller, R. Koch. Pose estimation and map building with a Time of Flight camera for robot navigation. International Journal of Intelligence Systems Applications, vol. 5, pp. 355-364, 2008.
18. P. Pudil, F.J. Ferri, J. Novovicova, J. Kittler. Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions. Proceedings of 12th IAPR Iternational Conference vol. 2, pp. 278-283, 1994.
19. A. W. Whitney. A direct method of nonparametric measurement selection. IEEE Transactions on Computers, vol.20, pp. 1100-1103, 1971.
20. J. Shi, J. Malik, Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Inteligence, Vol. 22, No. 8, pp. 888-905, 2000.
21. M. Gschwandtner; R. Kwitt; A. Uhl, BlenSor: Blender Sensor Simulation Toolbox. Proceedings of 7th International Symposium In Advances in Visual Computing. ISVC 2011, 2011.

Isolated Spanish Digit Recognition based on Audio-Visual Features

Gonzalo D. Sad, Lucas D. Terissi and Juan C. Gómez

Lab. for System Dynamics and Signal Processing, Universidad Nacional de Rosario, Argentina
CIFASIS-CONICET, Rosario, Argentina
{sad, terissi, gomez}@cifasis-conicet.gov.ar

Abstract. The performance of classical speech recognition techniques based on audio features is degraded in noisy environments. The inclusion of visual features related to mouth movements into the recognition process improves the performance of the system. This paper proposes an isolated word speech recognition system based on audio-visual features. The proposed system combines three classifiers based on audio, visual and audio-visual information, respectively. An audio-visual database composed by the utterances of the digits (in Spanish language) is employed to test the proposed system. The experimental results show a significant improvement on the recognition rates through a wide range of signal-to-noise ratios.

Keywords: Speech recognition, audio-visual speech features, Hidden Markov Models

1 Introduction

In recent years, significant research efforts have been devoted to the development of Multimodal Human Computer Interfaces (HCIs) that try to imitate the way humans communicate with each other, which is inherently a multimodal process, in the sense that, for the transmission of an idea, not only is important the acoustic signal but also the facial expressions and body gestures [4]. For instance, a significant role in spoken language communication is played by lip reading. This is essential for the hearing-impaired people, and is also important for normal listeners in noisy environments to improve the intelligibility of the speech signal. Audio Visual Speech Recognition (AVSR) is a fundamental task in HCIs, where the acoustic and visual information (mouth movements, facial gestures, etc.) during speech are taken into account. Several strategies have been proposed in the literature for AVSR [7][6], where improvements of the recognition rates are achieved by fusing audio and visual features related to speech. As it is expected, these improvements are more notorious when the audio channel is corrupted by noise, which is a usual situation in speech recognition applications. These strategies usually differ in the way the audio and visual information is extracted and

combined, and the AV-Model employed to represent the audio-visual information. These approaches are usually classified according to the method employed to combine (or fuse) the audio and visual information, *viz.*, feature level fusion, classifier level fusion and decision level fusion [2].

In feature level fusion (early integration), audio and visual features are combined to form a unique audio-visual feature vector, which is then employed for the classification task. This strategy is effective when the combined modalities are correlated, since it can exploit the covariance between the audio and video features. This method requires the audio and visual features to be exactly at the same rate and in synchrony, and usually performs a dimensionality reduction stage, in order to avoid large dimensionality of the resulting feature vectors. In the case of classifier level fusion (intermediate integration), the information is combined within the classifier using separated audio and visual streams, in order to generate a composite classifier to process the individual data streams [5]. This strategy has the advantage of being able to handle possible asynchrony between audio and visual features. In decision level fusion (late integration), independent classifiers are used for each modality and the final decision is computed by the combination of the likelihood scores associated to each classifier [3]. Typically, these scores are fused using a weighting scheme defined based on the reliability of each unimodal stream. This strategy does not require strictly synchronized streams.

In this paper an isolated digit recognition system based on audio-visual features is proposed. This system is based on the combination of early and late fusion schemes. In particular, acoustic information is represented by mel-frequency cepstral coefficients, and visual information is represented by coefficients related to mouth shape. The efficiency of the system is evaluated considering noisy conditions in the acoustic channel. The proposed system combines three classifiers based on audio, visual and audio-visual information, respectively, in order to improve the recognition rates through a wide range of signal-to-noise ratios (SNRs). A Spanish audio-visual database is employed to test the proposed system. The experimental results show that a significant improvement is achieved when the visual information is considered.

The rest of this paper is organized as follows, the audio, visual and audio-visual features used for each classifier are described in section 2 together with the database used for the experiments. The proposed classifiers and the early integration strategy are analyzed in section 3. A general description of the proposed system using the late fusion scheme is given in section 4. In section 5 experimental results are presented, where the performance of the proposed strategy is analyzed. Finally, some concluding remarks and perspectives for future work are included in section 6.

2 Audiovisual Database and Features

In order to evaluate the proposed speech recognition system an audio-visual database was compiled. This database consists of videos of 16 speakers facing

the camera, pronouncing a set of ten words 20 times, in random order. These words correspond to the Spanish utterances of the digits from zero to nine. The videos were recorded at a rate of 60 frames per second with a resolution of 640×480 pixels, and the audio was recorded at 8 kHz synchronized with the video. All the recorded words in the videos were automatically segmented based on the audio signal, by detecting zero-crossings and energy level in a frame wise basis.

The audio signal is partitioned in frames with the same rate as the video frame rate (60 frames per seconds). For a given frame t, the first eleven non-DC Mel-Cepstral coefficients are computed and used to compose a vector denoted as $\mathbf{a}_t$. In order to take into account the audio-visual co-articulation, the information of t_{c_a} preceding and t_{c_a} subsequent frames is used to form the audio feature vector at frame t, $\mathbf{o}_{at} = \left[\mathbf{a}_{t-t_{c_a}}, \ldots, \mathbf{a}_t, \ldots, \mathbf{a}_{t+t_{c_a}}\right]$.

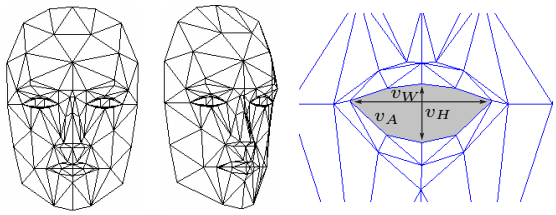
Visual features are represented in terms of a simple 3D face model, namely *Candide-3* [1]. This 3D face model, depicted in Fig. 1(a), has been widely used in computer graphics, computer vision and model-based image-coding applications. The advantage of using the Candide-3 model is that it is a simple generic 3D face model, adaptable to different real faces, that allows to represent facial movements with a small number of parameters. The method proposed by the present authors in [8] is used to extract visual features related to mouth movements during speech. As it is described in [8], this visual information is related to the generic 3D model and it does not depend on the particular face being tracked, *i.e*, this method retrieves normalized mouth movements. The mouth shape at each frame t is then used to compute three visual parameters, *viz.*, mouth height (v_H), mouth width (v_W) and area between lips (v_A), as depicted in Fig. 1(b). These three parameters are used to represent the visual information at frame t, denoted as $\mathbf{v}_t$. Similarly to the case of acoustic information, t_{c_v} preceding and t_{c_v} subsequent frames are used to form the visual feature vector at frame t, $\mathbf{o}_{vt} = \left[\mathbf{v}_{t-t_{c_v}}, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_{t+t_{c_v}}\right]$.

For a particular frame t, the audio-visual feature vector is composed by the concatenation of the associated acoustic and visual feature vectors, that is

$$\mathbf{o}_{avt} = \left[\mathbf{o}_{at}, \mathbf{o}_{vt}\right]. \tag{1}$$

3 Early Integration

In most applications the acoustic channel is corrupted by noise, degrading the recognition rates of audio-only speech recognition systems. The proposed system aims to improve the recognition rates in these situations, by fusing audio and visual features. In the presence of noise in the acoustic channel, the efficiency of a classifier based on audio-only information decreases as the SNR decreases. On the other hand, the efficiency of a visual information classifier remains constant, since it does not depends on SNR. However, the use of only visual information is usually not enough to obtain relatively good recognition rates. It has been shown in several works in the literature [4][7][6], that the use of audio-visual



(a) (b)

Fig. 1. (a) *Candide-3* face model. (b) Visual parameters.

feature vectors (early integration) improves the recognition rate in the presence of noise in comparison to the audio-only case. In this section, the performances of audio, visual, and audio-visual classifiers are evaluated using audio-visual features extracted from the compiled database, described in section 2. Then, these results are used to derive the proposed late integration strategy described in section 4.

Visual classifier. The visual feature vector $\mathbf{o}_{vt}$ at frame t is composed by the concatenation of the visual information contained in t_{c_v} preceding and t_{c_v} subsequent frames (see section 2). Experiments with 0 to 7 frames of coarticulation (t_{c_v}) were carried out. It must be noted that there is no need to carry out these tests considering different SNRs, since the visual features are not affected by the acoustic noise. The results of these experiments are depicted in Fig. 2, using boxplot representation. As it is customary, the top and bottom of each box are the 75th and 25th percentiles of the samples, respectively, the line inside each box is the sample median, and the notches display the variability of the median between samples. These results were computed across all the words in the vocabulary.

Audio classifier. Similarly to the case of visual feature vectors, the audio feature vector $\mathbf{o}_{at}$ at frame t is composed by the concatenation of the acoustic information contained in t_{c_a} preceding and t_{c_a} subsequent frames. To select the optimum value of t_{c_a}, experiments with 0 to 6 frames of coarticulation were performed. Since the efficiency of the audio classifier depends on the SNR, these experiments were carried out using several SNR levels for two types of noise: additive Gaussian noise and Babble noise. In Fig. 3, the results derived from these experiments are shown, where only the medians for each noise level and coarticulation parameter, are depicted for visual clarity reasons.

Audio-Visual classifier. The audio-visual fusion (early integration) proposed in this paper is based on the concatenation of the audio and visual feature vectors associated to each frame t, as stated in (1). Thus, there are two parameters that define the audio-visual vector: t_{c_a} and t_{c_v}. Modifying these values

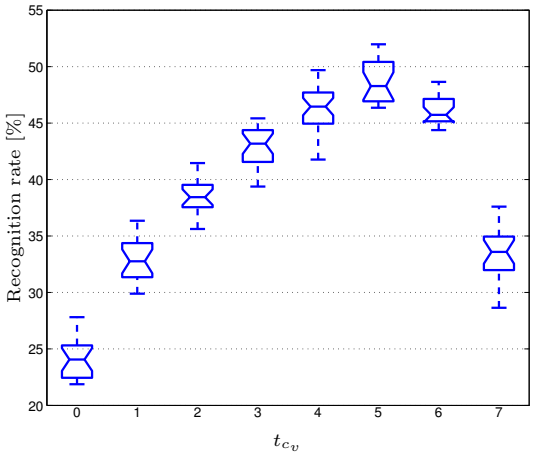


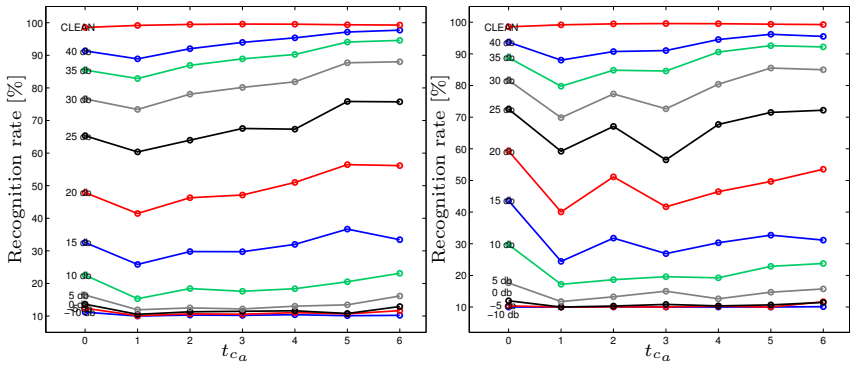
Fig. 2. Recognition rate of the visual classifier using different values of t_{c_v}.

different structures can be obtained. In a similar fashion than for the audio classifier and video classifier, experiments were performed for t_{c_a} and t_{c_v} ranging from 0 to 5. These tests were carried out considering different SNRs for the cases of Gaussian and Babble noises. Figure 4 shows the recognition rates obtained for the different SNRs and the two considered noises, for three particular audio-visual fusion configurations, namely

- $t_{c_a} = 1$ and $t_{c_v} = 5$, denoted as A_1V_5,
- $t_{c_a} = 5$ and $t_{c_v} = 5$, denoted as A_5V_5,
- $t_{c_a} = 5$ and $t_{c_v} = 1$, denoted as A_5V_1.

It can be noted from Fig. 4 that the better performance at low SNRs is obtained for the case of configuration A_1V_5, while configurations A_5V_5 and A_5V_1 present the better performances at high SNRs. The performance of the remaining configurations lies between these curves following the same properties.

Considering the results associated to each classifier, depicted in Figures 2, 3 and 4, it can be clearly noted that the audio classifier performs better than the visual one for high SNRs and viceversa. The combination of audio-visual features leads to an improvement of the recognition rates in comparison to the audio-only case. However, for the case of low SNRs, the audio-visual classifier performs worse than the visual one since fused audio-visual features are degraded by the highly corrupted acoustic data. Using different combination of acoustic and visual features, different performances can be obtained. For instance, if the audio-visual features contains more visual than acoustic information, the performance at low SNRs is improved since visual information is more reliable in this case. However, the efficiency at high SNRs is deteriorated, where the acoustic information is more important. Even for cases where a small portion of



(a) Babble Noise (b) Gaussian noise

Fig. 3. Efficiency of the acoustic classifier using different values of t_{c_a} and different SNRs, for the cases of considering (a) Babble noise, and (b) Gaussian noise.

audio information is considered, a notorious improvement could be obtained for low SNRs, but the efficiency at high SNRs could be worse than for the audio-only case. Thus, there exists a trade-off between performance at low and high SNRs.

4 Late Integration

Taking into account the analysis presented in the previous section, the recognition system proposed in this paper combines three different classifiers based on audio, visual and audio-visual information, respectively, aiming at recognizing the input word and maximizing the efficiency over the different SNRs. In the training stage, a combined classifier is trained for each particular word in the vocabulary. Then, given an audio-visual observation sequence associated to the input word to be recognized, denoted as O_{av}, which can be partitioned into acoustic and visual parts, denoted as O_a and O_v, respectively, the probability (P_i) of the proposed combined classifier corresponding to the i-class is given by

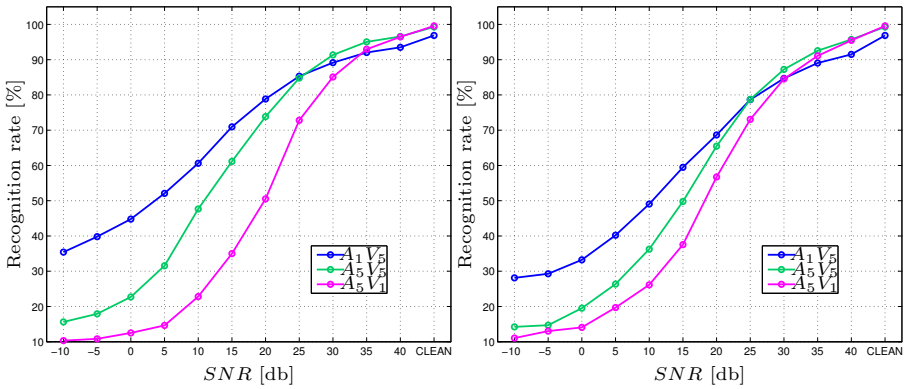
$$P_i = P\left(O_a|\lambda_i^a\right)^{\alpha} P\left(O_v|\lambda_i^v\right)^{\beta} P\left(O_{av}|\lambda_i^{av}\right)^{\gamma}, \tag{2}$$

where $P\left(O_a|\lambda_i^a\right)$, $P\left(O_v|\lambda_i^v\right)$ and $P\left(O_{av}|\lambda_i^{av}\right)$ are the probabilities corresponding to the audio (λ_i^a), visual (λ_i^v) and audio-visual (λ_i^{av}) classifiers, respectively, and α, β and γ are real coefficients that satisfy the following condition

$$\alpha + \beta + \gamma = 1. \tag{3}$$

The visual (λ_v) classifier is more useful at low SNRs (β is predominant), where the acoustic data is highly corrupted by noise, while at medium levels of SNRs, the audio-visual classifier (λ_{av}) retrieves the better decisions (γ is predominant). For high SNR conditions, an audio classifier (λ_a) is employed (α is predominant).

A block diagram representing this computation is depicted in Fig. 5.



(a) Babble Noise (b) Gaussian noise

Fig. 4. Performance of the audio-visual classifier over the SNRs for three different fusion configurations. (a) Babble noise. (b) Gaussian noise.

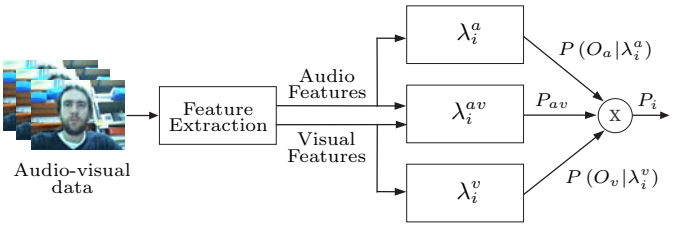


Fig. 5. Schematic representation of the computation of the probability associated to a particular class i for the proposed combined classifier. P_{av} refers to $P(O_{av}|\lambda_i^{av})$.

5 Experimental Results

In this section, the proposed audio-visual speech recognition system is evaluated using the audio-visual database described in section 2. For each experiment reported in this section, 50 round cross-validation was performed, randomly selecting 70% of the database for training and using the remaining 30% for testing. In these experiments the coefficients of the feature vectors were normalized by subtracting the corresponding sample mean and dividing by the corresponding sample variance, computed over the corresponding training set. To evaluate the recognition rates under noisy acoustic conditions, experiments with additive Gaussian noise and Babble noise, with SNRs ranging from -10 dB to 40 dB, were performed.

As it was previously described, the proposed audio-visual speech recognition system combines three classifiers based on audio, visual and audio-visual information, respectively, in order to improve the recognition rates for different SNRs.

These individual classifiers are implemented using left-to-right Hidden Markov Models (HMM) with continuous observations. In order to select the optimum HMM structure, several experiments were performed considering numbers of states in the range from 3 to 7, numbers of Gaussian mixtures from 4 to 11, and full and diagonal covariances matrices. These tests were carried out for the three cases, namely audio, visual and audio-visual features. Based on these experiments, an optimum HMM structure with 4 states, 6 Gaussian mixtures and full covariance matrices was selected for the three different classifiers.

5.1 Classifier selection

For the visual classifier, the results depicted in Fig. 2 shown that the higher accuracy was obtained for 5 frames of coarticulation, which corresponds to a visual feature vector $\mathbf{o}_{vt}$ composed by 33 parameters. In the time domain, this corresponds to a sliding window of 183 msec approximately. Thus, $t_{c_v} = 5$ was adopted for this classifier.

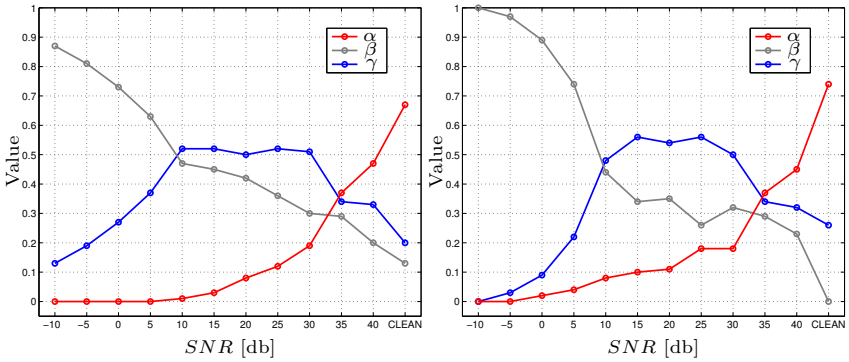
For the audio classifier, it must be noted that the selection of t_{c_a} should be done taking into account that the contribution of this classifier to the final decision stage is important at high SNR conditions. For that reason and looking at Fig. 3, $t_{c_a} = 4$ or $t_{c_a} = 5$ or $t_{c_a} = 6$ could be selected. In order to reduce the dimensionality of the resulting audio feature vectors, without significatively affecting the efficiency of the classifier, $t_{c_a} = 4$ was adopted, which corresponds to audio feature vectors composed by 99 parameters. In the time domain, this corresponds to a sliding window of 150 msec.

Regarding the selection of the optimal audio-visual classifier configuration to be used at the final decision stage, it must be taken into account that the contribution of this classifier is important at low and middle range SNR conditions, since at high SNR the audio classifier provides more accurate decisions. Thus, from Fig. 4 an adequate configuration for this purpose is the one using $t_{c_a} = 1$ and $t_{c_v} = 5$, *i.e.*, configuration A_1V_5.

5.2 Decision level integration

As mentioned in section 4, the combination of the probabilities computed from the independent classifiers, is carried out by the weighted multiplication of the individual probabilities, see Eq. (2), where coefficients α, β and γ modify the contribution to the final decision of the audio, visual and audio-visual classifiers, respectively. The values of these coefficients should be modified for the different SNRs, so that the higher contribution at low SNR comes from the visual classifier, at medium SNRs from the audio-visual classifier, and at high SNRs from the audio classifier. Several experiments were performed using different possible combinations of them to achieve the optimum values. The results of these test are depicted in Fig. 6. For both cases of considering Gaussian and Babble noises, it can be seen that the optimum value of α is the lower one at low SNRs, and it increases as the SNR increases, becoming the higher one at clean audio. On

the other hand, the optimum values of coefficient β present an inverse evolution. While for the case of coefficient γ the higher values are at medium SNRs.



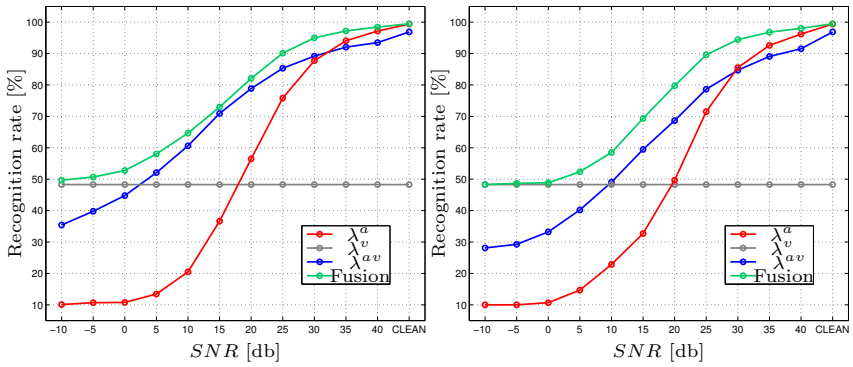
(a) Babble Noise (b) Gaussian noise

Fig. 6. Optimum values for coefficients α, β and γ over the SNRs. (a) Babble noise. (b) Gaussian noise.

In Fig. 7 the obtained recognition rates of the proposed fusing strategy over the SNRs, using the optimum values for the weighting coefficients α, β and γ, are presented. In this figure, the recognition rates corresponding to the audio, visual and audio-visual classifiers are also depicted. It is clear that the proposed objective of improving the recognition rates through the different SNRs has been accomplished.

6 Conclusions

Improvements of speech recognition rates by the incorporation of visual data related to the mouth movements and the late integration of different classifiers are presented in this paper. An isolated Spanish digit recognition system based on audio-visual information was developed to test the proposed system. The acoustic information is represented by mel-frequency cepstral coefficients, while the visual information is represented by coefficients related to mouth shape. Three classifiers based on audio, visual and audio-visual information, respectively, are combined in the proposed system in order to improve the recognition rates through a wide range of signal-to-noise ratios. A Spanish audio-visual database was compiled in order to evaluate the efficiency of the system, considering noisy conditions in the acoustic channel. The experimental results show that a significant improvement is achieved when the visual information is considered. It is important to note that, the absolute recognition rates could be further improved by considering well-known strategies usually employed in speech recognition, for instance, using delta mel-cepstral coefficients in the audio features,



(a) Babble Noise (b) Gaussian noise

Fig. 7. Recognition rates of the proposed fusing strategy, audio, visual and audio-visual classifiers.

including noisy features in the training stage, etc. Work is in progress, where the extension of the proposed system to the case of continuous speech recognition is considered.

References

1. Ahlberg, J.: Candide-3 - an updated parameterised face. Tech. rep., Department of Electrical Engineering, Linkping University, Sweden (2001)
2. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia 2(3), 141–151 (Sep 2000)
3. Estellers, V., Gurban, M., Thiran, J.: On dynamic stream weighting for audio-visual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(4), 1145–1157 (2012)
4. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: A survey. Comput. Vis. Image Understand 108(1-2), 116–134 (2007)
5. Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A coupled hmm for audio-visual speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (CASSP02). pp. 2013–2016 (2002)
6. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent advances in the automatic recognition of audio-visual speech. In: PROC. IEEE. vol. 91, pp. 1306–1326 (2003)
7. Shivappa, S., Trivedi, M., Rao, B.: Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. Proceedings of the IEEE 98(10), 1692–1715 (2010)
8. Terissi, L., Gómez, J.: 3D head pose and facial expression tracking using a single camera. Journal of Universal Computer Science 16(6), 903–920 (2010)

Desarrollo de una Ficha Anestésica Web en Áreas criticas

Gustavo Bianco [21], Marcelo Sabalza[1], Daniel Luna[1],Gustavo García Fornari[2], Jorge Garbino[1], Martín Waldhorn[2], Estefania Tarsetti[1]

[1] Departamento de Informatica en Salud, Hospital Italiano de Buenos Aires, Juan D. Peron 4190, Capital Federal, Argentina

[2] Servicio de Anestesia, Hospital Italiano de Buenos Aires, Juan D. Peron 4190, Capital Federal, Argentina

E-mail: gustavo.bianco@hospitalitaliano.com.ar;

Abstract. Este trabajo se centra en el diseño e implementación de un sistema de registro anestésico web en tiempo real del cual se genera un documento de relevancia asistencial y legal. La solución abarca un híbrido de una aplicación web integrada en la historia clínica electrónica y una aplicación local que maneja la comunicación con el monitor de signos vitales. Debido a la criticidad del ámbito de trabajo se buscó que pueda funcionar en contingencia, logrando una aplicación robusta y confiable..

Palabras clave: Signos Vitales. Anestesiología. Ficha anestésica. Historia clínica. Informática. Tiempo real.

1 Introducción

La ficha o registro anestésico es la documentación escrita y gráfica de lo que sucede durante un procedimiento anestésico. Es un documento que cumple fines médicos, legales, de investigación, docentes, estadísticos/epidemiológicos y de referencia para la facturación. A pesar de la importancia del registro anestésico, este tiene un rol secundario dentro del quirófano, ya que la prioridad del anestesiólogo es atender al paciente [1]. El llenado de este registro en papel es manual en la mayoría de las instituciones y los signos vitales como la frecuencia cardíaca, saturación de oxígeno, concentración de dióxido de carbono en la vía aérea, temperatura y tantos otros tienen que registrarse con una frecuencia mínima de 5 minutos [2]. Se ha demostrado que sin entrenamiento previo y con una inversión de bajo costo, se puede implementar en las salas de operaciones el manejo automático de la información anestésica, teniendo siempre presente que para la justicia, una buena ficha anestésica presupone siempre una buena praxis [2].

La bibliografía reporta que en Estados Unidos en 1998 apenas el 1% de los departamentos de anestesia utilizaban sistemas de documentación informáticos en la sala de cirugía, y se estima que actualmente menos del 10 % de todos los hospitales cuentan con este tipo de sistemas [3]. Un estudio demostró que el 70% de los incidentes que ocurren durante el proceso de anestesia están relacionados a errores humanos, y algunos de estos incidentes muestran una falla de la comunicación funcional entre el personal médico.[4] La exactitud de la gráfica anestésica tradicional parece reducirse además significativamente en caso de incidentes críticos. Por ejemplo se observó que más del 22% de los valores registrados por 10 anestesiólogos sometidos a un incidente crítico complejo simulado, anotaron valores que discrepaban en más de un 25% de la realidad, e incluso se registraron errores superiores al 100% de la realidad. Otro aspecto importante es la posibilidad de realizar análisis posteriores por ejemplo; Benson y Col. revisaron 16.019 anestesias para localizar la existencia de episodios de hiper o hipotensión arterial, bradicardia, taquicardia e hipovolemia. Estos fueron recogidos en 911 pacientes (5,7%) de forma manual y en 2.996 pacientes (18,7%) de modo automatizado [5].

Se investigó desarrollos de software que asisten en la tarea de completar el registro electrónico con captura automática de signos vitales en tiempo real, por ejemplo MV-OR de iMDsoft[6], SAFERsleep de la empresa del mismo nombre[7] y CompuRecord de Philips [8]. Se vieron varias alternativas de arquitecturas y diseños de interfaces prestando especial atención a este último punto y a la usabilidad. La mayoría del software comercial que realiza registro anestésico son aplicaciones de escritorio, las que no coinciden con el lenguaje de programación y los criterios de ubicuidad, accesibilidad y alta disponibilidad de las tecnologías de desarrollo del Departamento de Informática en salud del Hospital Italiano de Buenos Aires (HIBA). Esta institución cuenta con un sistema de salud informatizado donde la Historia Clínica Electrónica (HCE) de desarrollo propio es su aplicación central y es el repositorio de la documentación de todo acto médico [9]. La HCE es una aplicación web y se busca que la mayoría de las aplicaciones desarrolladas para interoperar con la misma también lo sean. Siendo el registro anestésico una actividad que se desarrolla en muchos casos dentro de los quirófanos, las aplicaciones web ubicadas en servidores centralizados y dependientes de redes de comunicación fisicamente distribuídas (intranets) entran en conflicto con la normativa para sistemas médicos en áreas críticas (IEC60601), las cuales exigen que los sistemas que allí se utilicen se encuentren aislados de las redes eléctricas y de las redes de datos externas.

El desarrollo de la Ficha Anestésica Electrónica (FAE) surge como respuesta a múltiples problemáticas, algunas generales y otras particulares de esta Institución. Entre las generales se puede nombrar:

- Problemas derivados de registros en papel, se pierden en el traslado, se traspapelan o son ilegibles, esto genera problemas médicos, legales y de facturación.
- La forma de digitalizarlos es realizando un escaneo (siendo muy complicado y costoso de utilizar lo registrado para estadísticas).
- Los registros se deben almacenar por ley un mínimo 10 años (SALUD PÚBLICA - Ley 26.529).

- Se estima que entre el 10 y 15 % del tiempo del anestesiólogo se utiliza para completar la ficha anestésica convencional, esta distracción implica un alto riesgo para el paciente [3].

Particularmente del HIBA:

- El creciente número de quirófanos como así también de las cirugías que allí se realizan genera una cantidad creciente de información a almacenar.
- Dado que el HIBA posee un sistema de Historia Clínica Electrónica, la ficha convencional queda fuera del sistema informático.
- Se dificulta cualquier tipo de análisis estadístico o de investigación.
- El HIBA tiene la Iniciativa papel cero.

El desarrollo de la FAE busca resolver a largo plazo los problemas anteriormente enunciados como así también incluir toda la información que se registra actualmente en la ficha convencional, automatizando la mayor cantidad de tareas posibles. En este trabajo se describe el desarrollo de una solución de software y hardware para el registro en línea de una ficha anestésica electrónica web con captura automática en tiempo real de signos vitales en áreas críticas, respetando las normativas vigentes y los criterios propios de la institución.

2 Materiales y Métodos

El Hospital Italiano es un Hospital Universitario de alta complejidad fundado en 1853, pertenece a una red sanitaria sin fines de lucro que incluye 2 hospitales, 23 centros periféricos ambulatorios y 150 consultorios particulares. En la red trabajan 2500 médicos, 1000 enfermeros y 2500 profesionales del equipo de salud provenientes de otras disciplinas. Con el apoyo de 1500 administrativos, los profesionales atienden 2.800.000 consultas ambulatorias y 50.000 internaciones anuales que se distribuyen en sus 750 camas (200 de cuidado críticos). Desde 1998 el HIBA cuenta con un sistema de información en salud integrado. Su historia clínica electrónica (HCE) es web, desarrollada en Java y es el repositorio de la documentación de todo acto médico.

Desde el 2013 el HIBA cuenta con 30 quirófanos donde se desempeñan los anestesiólogos, de los cuales 15 pertenecen al Quirófano Central (QC). Existen normas como la IEC60601 que exigen para la atención segura del paciente que las redes de datos y eléctricas deben permanecer aisladas del entorno externo al quirófano durante la cirugía. En la etapa inicial se aprovechó que los nuevos quirófanos centrales estaban en construcción para poder preparar una infraestructura acorde a las necesidades de la FAE. Se diseñó de tal forma que cada quirófano cuenta con un rack informático propio. Además se planteó que cada quirófano sea una unidad funcional independiente y que la FAE cumpla con los siguientes requerimientos:

- Tiene que poder seguir funcionando ante la eventual falla de los servidores y/o conexiones externas.

- En el caso de perder la conexión con los servidores externos, al restablecerse esta conexión debe sincronizar todo los datos que hayan quedado pendientes de actualizar.
- Debe considerar un funcionamiento en contingencia.
- Debe ser ejecutada desde dentro de la historia clínica del paciente.
- Debe ser Web y en el mismo lenguaje que la HCE (JAVA).

2.1 Hardware

Para la primera etapa, en el QC, se diseñó una solución dividida en 2 partes:

- Dentro del quirófano se ubicó un Monitor Touch de grado médico fijado mediante un brazo metálico articulado a la mesa de anestesia y un soporte porta teclado y mouse.
- En el rack contiguo se alojó un CPU y la comunicación entre ambos se realiza mediante un bloque modulador/demodulador.

Este diseño se justifica en que el CPU no es un equipo de grado médico y por lo cual no puede estar alojado dentro del quirófano. Como la distancia entre el CPU y los dispositivos de interfaz humana es de más de 20 metros fue necesario un bloque que module las señales para transmitirlas entre los dispositivos.

Los datos son obtenidos de monitores multiparamétricos Philips modelos MP y MX ubicados en las mesas de anestesia.

2.2 Software

Se optó por una arquitectura dividida en tres partes: la interfaz de usuario, la interfaz con la HCE y la interfaz con el monitor de signos vitales. A continuación se pude ver un esquema arquitectura propuesta y de sus interrelaciones (Figura 1).

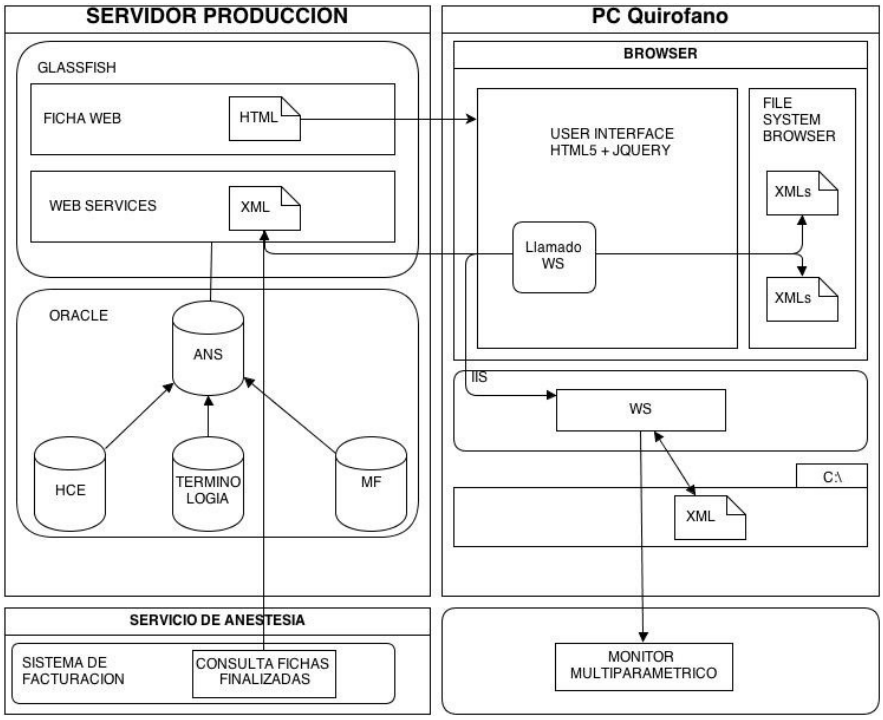


Figura 1: Diagrama de Arquitectura.

La interfaz con el monitor multiparamétrico es un servicio web (webservice) instalado en el IIS del CPU en el rack del quirófano que está físicamente conectado al monitor de Signos Vitales. El webservice fue programado en .net como evolución de una aplicación pre-existente (SVCaptor) de desarrollo propio y utilizada desde hace 3 años de forma rutinaria en las terapias y servicios de emergencia del hospital para el registro automático de signos vitales en la HCE desde los monitores paciente Philips [10] . El protocolo de comunicación subyacente fue implementado sobre la capa de transporte física Ethernet y de comunicación lógica del protocolo UDP/IP en base a la interfaz de exportación de datos que tienen los monitores Philips, serie MP y MX y que se ajusta con bastante fidelidad al protocolo estándar de comunicación en tiempo real ISO/IEEE 11073, En este contexto, el CPU de quirófano será el cliente y el monitor el servidor (Figura 2).

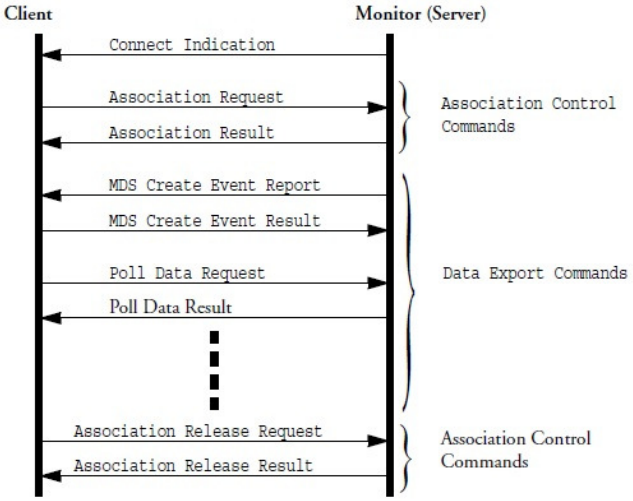


Figura 2: Diagrama del protocolo de comunicación.

El monitor se configura para que entregue en tiempo real (una muestra por segundo) los signos vitales que le está ingresando a través de sus sensores. Una vez establecida la conexión con el monitor lo siguiente es recibir y filtrar las tramas Ethernet con los paquetes UDP portadores de los mensajes con la información de los signos vitales, parsear estos mensajes, identificar los signos vitales recibidos y tomar correctamente los valores de cada uno. Como el volumen de información es considerable se tomó la decisión de tomar una muestra por minuto de cada signo vital y almacenarla a disco localmente, dejando la potencialidad de configurar este parámetro a futuro.

El web service realiza la interfaz entre la aplicación web y el monitor exponiendo métodos o funciones que le permiten iniciar la conexión, informar signos vitales disponibles, configurar cuales se desean capturar, comenzar la captura e informarlos.

La interfaz con la HCE recoge de las bases de datos las cirugías programadas con los datos de los pacientes, de los médicos y ofrece una serie de servicios web llamados desde la interfaz del usuario. También al momento de tener que sincronizar la ficha anestésica electrónica se comunica con estos servicios. Una vez que se firma digitalmente la ficha se ejecutan tareas para integrar la información médica (evoluciones, problemas, prácticas, medicamentos, etc) con la HCE.

La interfaz de usuario se basa en una aplicación web HTML utilizando las ventajas de los estándares de HTML5 de caché de aplicaciones HTML y de Filesystem, pudiendo con estos APIs cachear la aplicación y los datos (tantos los de la base de datos como los generados desde la aplicación).

Con esta arquitectura logramos que al ejecutarse la aplicación se bajen a cache persistivo del browser los datos necesarios para completar la ficha anestésica electrónica durante el procedimiento anestésico y pudiendo grabar localmente en este cache los datos registrados durante la anestesia, corriendo como una aplicación offline con posibilidades de sincronizar con el server los datos recaudados.

Esta arquitectura es especialmente robusta ante la pérdida de conexión o caída de los servicios en los servidores de la institución y toleraría un downtime de 48 horas, ya que este es el tiempo de programación de quirófanos en la institución. En el caso

del que downtime sea mayor o sea un paciente de urgencia la aplicación puede funcionar sin los datos del paciente programado, cargando manualmente los datos relevantes del paciente y el episodio, para que una vez en funcionamiento la conexión con los servidores adjuntar la ficha anestésica electrónica a la historia clínica del paciente.

2.3 Registro convencional vs Electrónico

Se revisaron registros anestésicos convencionales en papel, detectando fácilmente sus puntos débiles; son de difícil lectura, dependen de la claridad del anestesiólogo y están acotadas en espacio. Algo que surge a posteriori es el desaprovechamiento de la información, ya que por su poca exactitud y confiabilidad no se utilizan para trabajos de investigación ni de análisis. Finalmente se realizó una observación en quirófano, identificando en qué momentos el anestesiólogo realiza el registro anestésico en papel y el tiempo que este conlleva. A partir de todo esto, los integrantes del equipo trabajaron sobre conceptos de usabilidad, diseñaron prototipos y maquetados de la interfaz táctil para testear casos de uso con los anestesiólogos. Esta instancia permitió incluir en la etapa de diseño y desarrollo a los usuario finales, los anestesiólogos, lo que se esperaba brinde un producto con mayor satisfacción y aceptación. Finalmente se elaboró una maqueta del documento que se genera al terminar la ficha anestésica electrónica y queda adosado a la historia clínica del paciente. De comparar el registro anestésico en papel escaneado y el nuevo documento que se genera la FAE se ve una gran diferencia en lo sencillo y legible que resulta entender lo registrado.

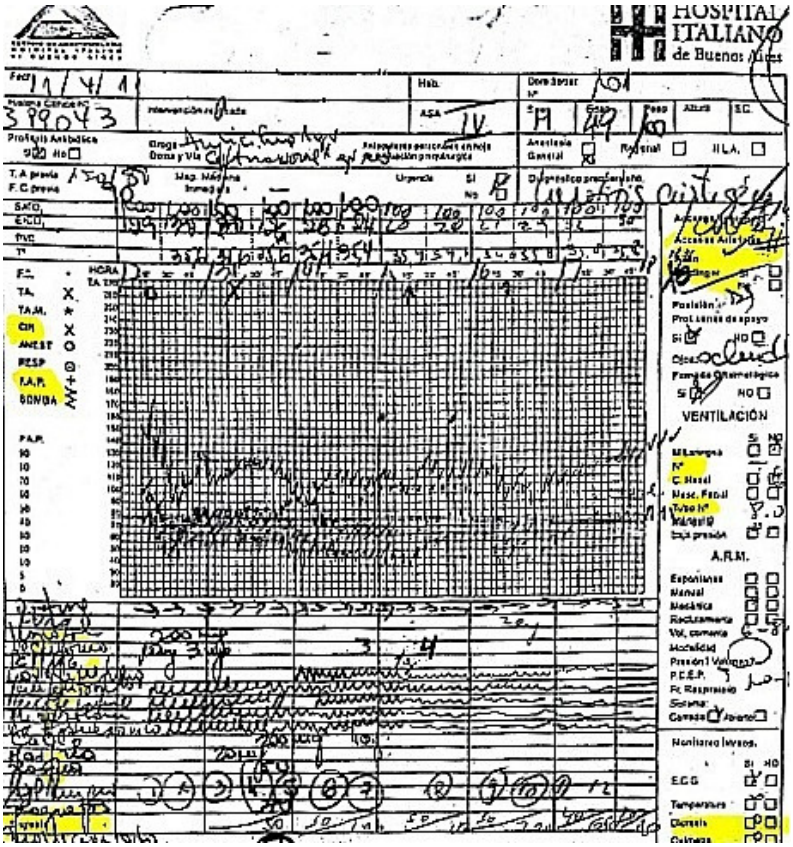


Figura 3: Ficha anestésica convencional.

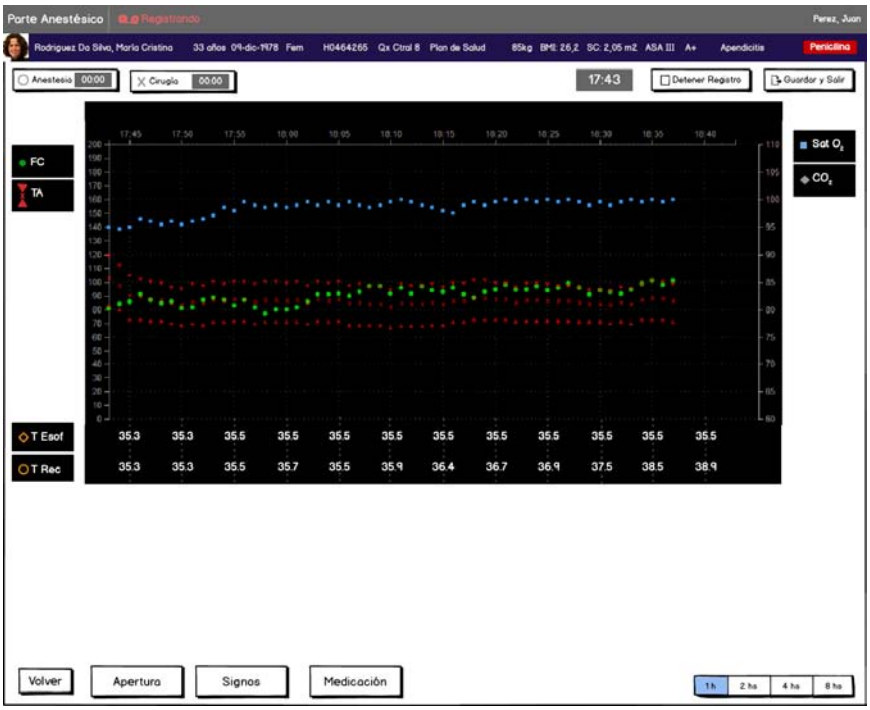


Figura 4: Captura de la Aplicación ejecutándose.

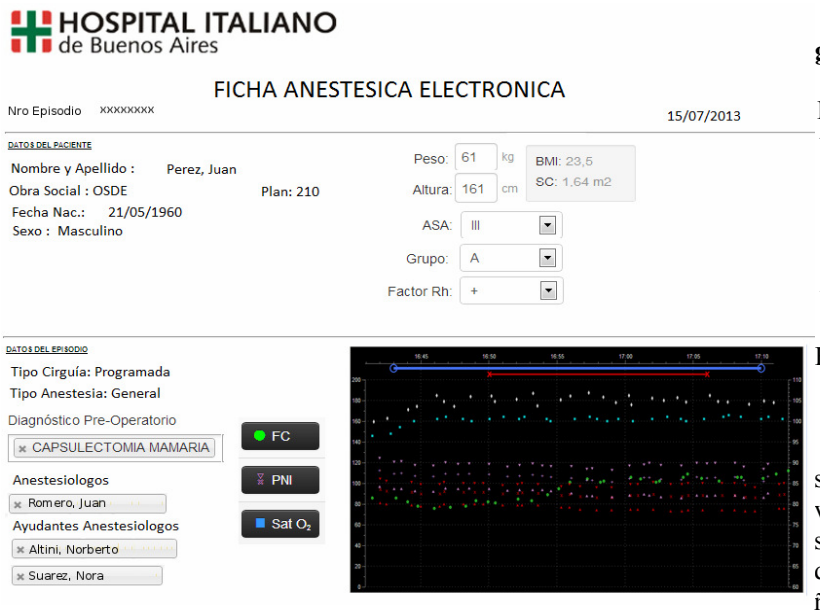


Figura 5: Maqueta del CDA que va a la HCE.

A su vez se diseñaro

n interfaces para interactuar con el sistema de facturación del servicio de anestesia donde la información llega directamente, sin riesgos de traspapelarse o perderse, siendo clara y ayudando al facturista a cargar correctamente las prácticas realizadas.

3 Resultados

En principio se logró cachear la información para que la aplicación pueda ejecutarse en el browser y no se vea afectada por la caída de los servidores de aplicaciones, bases de datos y/o fallas en la infraestructura física de la intranet. Luego se logró que la aplicación permita ejecutarse directamente en contingencia, sin haber logrado cachear los datos pertinentes antes de su ejecución y una vez restablecida la comunicación con el resto de los sistemas se sincronice con la base de datos central.

El webservice que se comunica con el monitor se puede configurar para que en lugar de almacenar 1 muestra por minuto pueda almacenar hasta 60 muestras por minuto. Esto está pensado para que en un futuro se pueda analizar que sucede con los signos vitales pocos segundos después de que se realiza una práctica o se administra una droga. Hay que tener en cuenta que la frecuencia de registro no se puede aumentar al máximo durante un periodo muy prolongado ya que los tamaños de los archivos crecen rápidamente. En esta primera etapa la aplicación solo maneja Signos vitales, si bien esto es suficiente para la ficha anestésica se busca también que a futuro se puedan almacenar tramos de señales como la de ECG, Presiones invasivas y/o concentración de gases.

Se realizaron 2 pruebas piloto en un quirófano del QC, primero se armó la parte de soporte que va unida a la mesa de anestesia, luego se realizó el montaje del CPU en rack contiguo y finalmente se colocó el monitor, mouse y teclado. Vale la pena resaltar que el brazo sobre el que va montado el monitor táctil permite regular su altura y posición para ajustarlo al usuario. Por otro lado el soporte donde van el teclado y mouse es rebatible ya que solo sería para situaciones excepcionales, pues normalmente todo se realizará desde la interfaz táctil.

Figura Y: Sistema de prueba ensamblado en quirófano.

Una vez montado el hardware se probó la aplicación con datos de un pacientes de prueba pero capturando y registrando los signos vitales de un paciente real. La idea de estas primeras pruebas era presentar la aplicación, probar como se navega desde la HCE hasta el acceso a la FAE y finalmente ver el desempeño de la interfaz gráfica y graficación de los signos vitales. Las pruebas dieron buenos resultados, los anestesiólogos pudieron interactuar fácilmente y de manera intuitiva, y la aceptación

de la aplicación fue alta. Si bien todos los anestesiólogos que probaron la aplicación tenían sugerencia para mejoras futuras, el 100% estaba conforme tanto con el Hardware como con el software.

Una vez finalizada la carga de la ficha anestésica electrónica, observamos que la información cargada representaba fielmente lo observado y capturado por el monitor, y se encontraba accesible para el resto de los sistemas (HCE, sistema de facturación, etc.).

4 Discusión

Si bien el desarrollo se encuentra en su primera etapa los resultados obtenidos son coherentes con la mayoría de la bibliografía al respecto. Desde el punto de vista de la integración con otros sistemas informáticos como ser los de Historia Clínica, Contables, etc. permite una comunicación instantánea, permitiendo reducir considerablemente los tiempos de espera y procesamiento que se tienen hoy en día con la ficha convencional. Otro aspecto que cambia de manera considerable es la confiabilidad y exactitud de la información que se genera por cirugía, permitiendo que en un futuro se puedan realizar trabajos de investigación con ella. A futuro no solo se planea implementar esta ficha en todos los quirófanos sino también en todos los sectores que se realicen procedimientos de anestesia, esto requerirá entre otras cosas, migrar la aplicación para que funcione en dispositivos móviles tipo tablet. Una vez completado el desarrollo se espera que la ficha no solo reemplace al papel, sino que agregue nuevas funcionalidades, como por ejemplo la posibilidad de que mediante inteligencia artificial se disparen alertas que asistan al anestesiólogo. Hay que tener en cuenta también que al digitalizarse la ficha anestésica esta contiene información sensible del paciente, por lo que se debe asegurar la seguridad y confidencialidad de la misma.

5 Conclusión

Esta primera etapa del desarrollo presenta mucho más que un saldo positivo, se logró establecer una conexión estable con el monitor Philips y realizar la captura de los signos vitales en tiempo real. Cabe destacar que el registro electrónico generado aumenta la frecuencia de muestreo 5 veces y asegura una confiabilidad del dato que la ficha convencional no posee. Se comprobó que la sincronización de datos previa y posteriormente al procedimiento anestésico permite un correcto funcionamiento en contingencia, lo cual le da una robustez y seguridad requerida para este tipo de aplicaciones. El grado de aceptación logrado fue muy satisfactorio teniendo en cuenta que se está introduciendo no sólo una aplicación, sino una nueva forma de hacer las cosas. En resumen, se logró desarrollar una solución informática a medida, partiendo de cero y con recursos propios que no solo cumple los requerimientos internos y normativa al respecto de la seguridad del paciente, sino que también contempla modernos conceptos de usabilidad y diseño, permitiendo una gran escabilidad y la posibilidad a futuro de uso en dispositivos móviles.

Bibliografía

1. Pini M., Lossetti O. and Trezza F., 2002 Suplemento del Diario del Mundo Hospitalario: *Importancia Medico-Legal de la Ficha Anestesica*, Año 6, Nº 26.
2. Capria J, Gómez Roca M., Tibaldi F., Wikinski J., 1997 Articulo de Investigación Clínica: *Comparación De La Información Obtenida Mediante Una Ficha Anestésica Manual Vs. Una Automática Computarizada*. 55: 03: 143-152.
3. Trivelato L., Pereira F., Smidarle D., Smidarle R. 2011, Revista Médica de Minas Gerais: *La Anestesiologia en la Era Digital*. ; Vol. 21(2 Supl 3): S28-S33.
4. Alapetite A., Gauthereau V. 2005 Annual Conference of the European Association of Cognitive Ergonomics: *Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room.* Section 2 pp. 189-196.
5. Ortiz-Gómez J. R., Monedero-Rodríguez P., Pérez-Cajaraville J. J. 2002 *Aplicaciones de la informática en Anestesiología: gráfica anestésica. 2002* 49: 141-149
6. Página Web de *iMDsoft* URL: http://www.imd-soft.com/mv-pacu
7. Página Web de *SAFERsleep* URL*:* http://www.safersleep.com
8. Página Web de *Philips* URL*: http://www.healthcare.philips.com/main/products/patient_monitoring/products/intellispace_cca/compurecord/*
9. Gonzalez Bernaldo de Quiros F, Soriano E, Luna D, Gomez A, Martinez M, Schpilberg M, Lopez Osornio, A. Desarrollo e implementación de una Historia Clínica Electrónica de Internación en un Hospital de alta complejidad. 6to Simposio de Informática en Salud - 32 JAIIO. Buenos Aires - Argentina – 2003
10. Bibiana Schachner, Antonio E. Arias, Jorge Garbino, Guillermo Vignau, Cintia Budalich, Daniel R. Luna, Fernán González B. de Quirós. Implementación de un Registro electrónico para Enfermería en una Unidad de Cuidados Intensivos del Adulto. Congreso Infolac - Guadalajara, Mexico. - 2011

Detección de signos respiratorios patológicos en poblaciones avícolas productivas mediante procesamiento digital de señales acústicas

Cristian Kühn and César Martínez[1,2]

[1] Laboratorio de Cibernética, Facultad de Ingeniería,
Universidad Nacional de Entre Ríos, Ruta 11, Km. 10, Oro Verde, Entre Ríos
[2] Centro de Investigación en Señales, Sistemas e Inteligencia Computacional
(SINC(i)), Dpto. Informática, Facultad de Ingeniería,
Universidad Nacional del Litoral, Santa Fe, Argentina
cristian.kuhn20@gmail.com,cmartinez@bioingenieria.edu.ar

Resumen La necesidad en detectar tempranamente la presencia de un problema sanitario en la producción avícola mejora sensiblemente las posibilidades para el control del mismo. Por ello, en éste trabajo se presenta el diseño y desarrollo de un método automático para la tarea de reconocimiento de signos patológicos respiratorios en forma temprana, orientado a la producción avícola. El sistema parte del registro de señales de pollos en galpones productivos, preprocesamiento para acondicionamiento de las señales, medición de parámetros de interés (energía, pseudoespectro) y generación de una señal de detección que indica la presencia de signos patológicos en la población estudiada. Los resultados obtenidos fueron satisfactorios, habiendo sido el sistema capaz de detectar los signos en diferentes condiciones de experimentación, desde el estudio de un solo individuo enfermo hasta la mezcla de individuos sanos y enfermos.

Keywords: análisis acústico, signos respiratorios patológicos, pseudoespectro, población avícola.

1. Introducción

La Industria Avícola es una de las cadenas productivas más importantes del país, habiéndose consolidado como una de las más dinámicas que tiene la producción agropecuaria. En esta industria, las enfermedades respiratorias de pollos son un tema de importancia sanitaria en un establecimiento productivo, dado que presentan una morbilidad alta (80-100 %) y la mortalidad oscila entre el 5 y el 20 %, según sea el tipo y severidad del brote [1].

Conocer la presencia de signos respiratorios en la población avícola es de gran importancia para tomar una acción temprana sobre el devenir en una enfermedad crónica. En la actualidad, y hasta el conocimiento de los autores, no se cuenta con un sistema confiable y de fácil aplicación que brinde este tipo de información. Uno de los principales problemas es la adquisición y clasificación *automáticas*,

ya que el registro audiovisual continuo es subjetivo, complicado y susceptible de errores.

El procesamiento digital de señales brinda herramientas que han sido aplicadas satisfactoriamente a diversas tareas, dando la posibilidad de contar con sistemas de implementación factible en el ambiente productivo animal. En este contexto, se han reportado diversas aplicaciones del análisis acústico estrechamente relacionadas con la aquí presentada, como ser el análisis de vocalizaciones de mamíferos [2], la comunicación de murciélagos [3], o el repertorio de sonidos de ballenas [4]. Una línea de trabajo previamente explorada por los autores consiste en el análisis acústico de sonidos masticatorios de rumiantes, a fin de automatizar el comportamiento ingestivo [5,6].

El estado del arte demuestra que el análisis acústico espectral resulta atractivo por su sencillez, rapidez y relativa robustez al ruido. Es por ello que en este trabajo se plantea el diseño y desarrollo de un método para la detección de patrones de signos respiratorios patológicos en sonido. Se trata de mantener una complejidad computacional relativamente baja, lo que brinda un sistema que ser utilizado para detectar en tiempo real la presencia de los signos anómalos en la producción avícola.

El resto del trabajo se organiza como se detalla a continuación. En la Sección 2 se expone detalladamente el diseño de la solución propuesta. En la Sección 3 se muestran los resultados obtenidos en diferentes condiciones experimentales. Finalmente, en la Sección 4 se resumen las conclusiones derivadas de este trabajo y se delinean trabajos futuros.

2. Método propuesto y materiales empleados

El problema de clasificación de signos respiratorios es similar a muchos problemas existentes en detección y clasificación de patrones. El proceso completo consta de las siguientes etapas, implementadas en el software matemático Matlab:

- El proceso de *recolección* de datos incluye la adquisición de audio usando dispositivos de grabación. Los registros son obtenidos en entornos controlados en cuanto al ruido de fondo.
- La *extracción de características* se basó en el examen espectro-temporal de los registros sonoros. Básicamente, consiste en el uso de mediciones de los picos principales en el pseudoespectro de segmentos de la señal. De la observación de los patrones se obtienen un grupo de parámetros que permitirán luego su discriminación.
- El *reconocimiento* consiste en la medición de los parámetros mencionados sobre intervalos de audio pseudo-estacionarios. El método considera una variabilidad permitida entre patrones, a fin de agregar robustez al sistema y ajustarse mejor a la realidad del problema.

La Figura 1 muestra un diagrama en bloques detallado de todo el proceso. A continuación se desarrolla cada uno de ellos, ejemplificando con las señales resultantes de cada proceso dada la novedad de la solución en la tarea abordada.

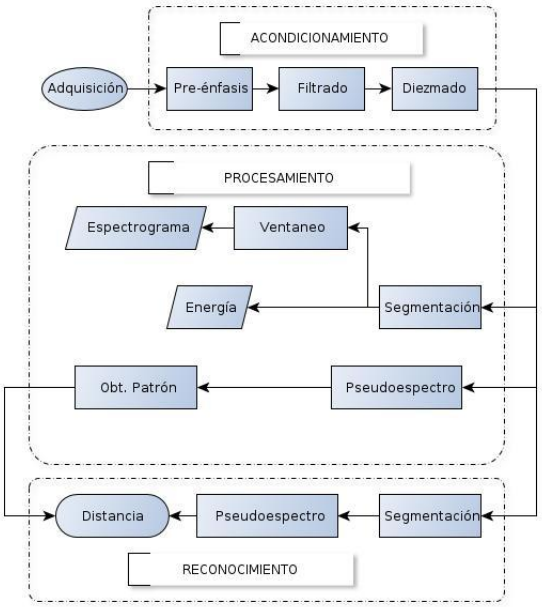


Figura 1. Diagrama completo del método propuesto para detección de signos respiratorios.

2.1. Adquisición y acondicionamiento

Adquisición. Se realizaron grabaciones digitales de 60 segundos de longitud conteniendo signos respiratorios patológicos de un grupo de aves afectadas. Las mismas fueron inicialmente apartadas de su entorno para minimizar los efectos del ruido de fondo. A priori se desconocían las características espectrales de los signos patológicos, por lo que se utilizó la resolución máxima disponible: 16 bits por muestra y una frecuencia de muestreo de 44100 Hz. Para las grabaciones se empleó el micrófono on-board de una Netbook *Asus I-EEE*.

La Figura 2 muestra un ejemplo de sonograma obtenido, de la cual se extrae el intervalo de tiempo entre los 20 y 30 s. a fin de evitar diversos ruidos: entre 10–30 s. se identifica el sonido de alboroto del pollo en tanto se estabilizaba frente al setup de adquisición, y entre 30–50 s. se registra la interferencia de un autómovil.

Pre–énfasis. La señal digitalizada $x(n)$ es sometida a un filtrado digital de bajo orden (típicamente, un filtro FIR de primer orden), para aplanar el espectro de la señal, según:

$$\widetilde{x}(n) = x(n) - ax(n-1); \text{ con } a \in \mathbb{R}.$$

Filtrado y diezmado. La implementación de esta etapa está dada por la necesidad de acotar la señal $\widetilde{x}(n)$ en banda, dejando pasar solamente aquéllas

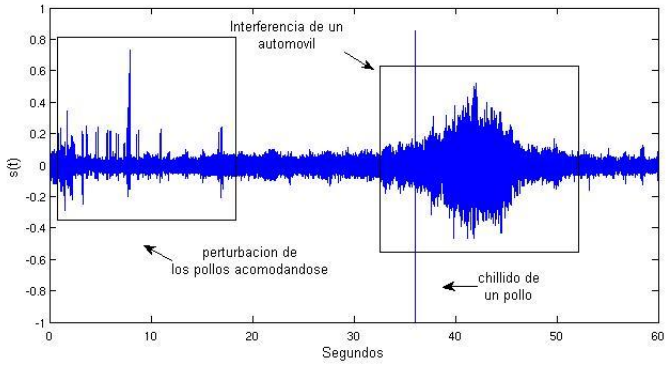


Figura 2. Sonograma de la pista de audio adquirida de 1 pollo con síntomas respiratorios.

frecuencias que contengan signos respiratorios. Para tal tarea se implementaron dos filtros Butterworth, un *pasa altos* y un *pasa bajos*, aplicados en ese orden respectivamente. La determinación de los parámetros de los filtros fue realizada experimentalmente por inspección del espectrograma de la señal, resultando así las frecuencias de cortes y paso necesarias. El diezmado es aplicado para submuestreo la señal por un factor entero en la forma $s(n) = \widetilde{x}(nK)$, preservando de la señal original una muestra de cada instante nK.

La Figura 2.1 muestra el sonograma filtrado y diezmado junto al espectrograma recortado en frecuencia a la banda de interés (0-2500 Hz). En ambos estudios se evidencian los patrones respiratorios patológicos periódicos (22 s., 24 s., etc.), inmersos en un ruido blanco de fondo.

2.2. Procesamiento de la señal

En este bloque se busca aislar los patrones de interés evidenciados en la señal. Para ello, se calculará el pseudoespectro de la señal y se determinará un patrón característico del signo respiratorio en la señal.

Para los procesos siguientes, la señal preprocesada $s(n)$ es ventaneada en bloques de N muestras con solapamiento del 50 %, con ventaneo de Hamming.

Energía. Una medida que ayuda a discernir los bloques con signos respiratorios es la energía de la señal, calculada como $\mathbf{E} = ||s(n)||_2^2$. La Figura 4 muestra un ejemplo de análisis, donde pueden observarse los picos en la localización de los eventos de interés.

Pseudoespectro. La estimación de las componentes frecuenciales de los signos respiratorios inmersos en la señal con ruido constituye la base para la clasificación. Así, se aplica en esta etapa el algoritmo *MUSIC (Multiple Signal Classification)* para la estimación del pseudoespectro de la señal acondicionada [7].

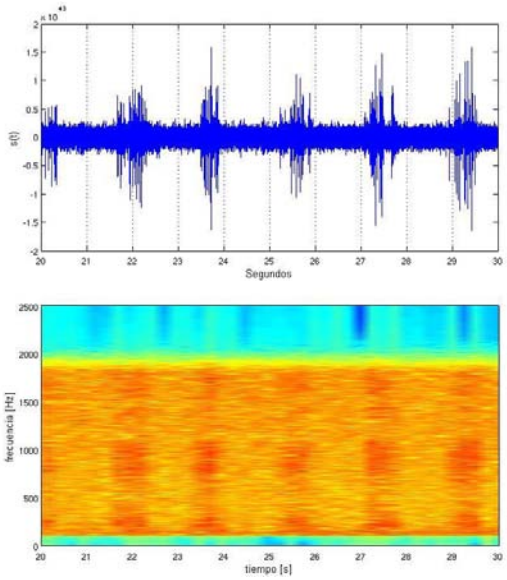


Figura 3. Sonograma y espectrograma de la señal $\widetilde{x}(n)$ luego de aplicarse los filtros pasa-altos y pasa-bajos, con posterior diezmado.

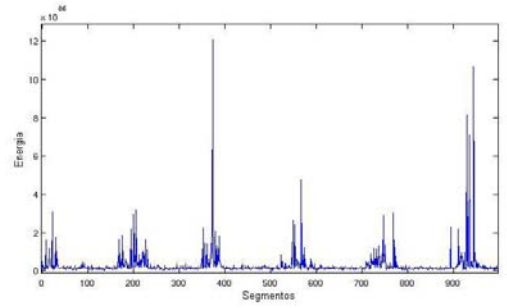


Figura 4. Energía contenida por cada bloque de la señal $s(n)$.

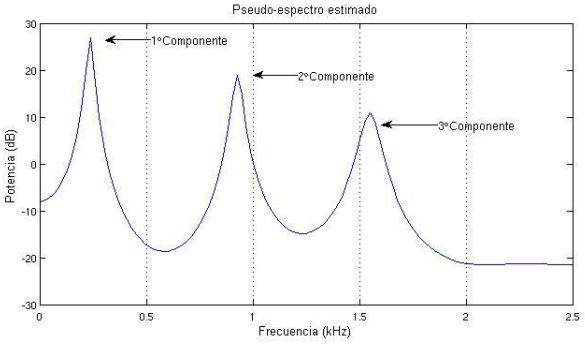


Figura 5. Pseudoespectro de la señal, con sus tres componentes de interés indicadas.

El algorithmo MUSIC logra estimar el contenido frecuencial de una señal pura contaminada con ruido blanco gaussiano, mediante una descomposición en valores y vectores propios, a lo que se denomina *pseudoespectro* [8]. La localización de los picos de la función estimada constituye la base de la detección de signos respiratorios en la señal. La Figura 5 muestra un ejemplo de pseudoespectro calculado sobre la señal de prueba utilizada.

2.3. Reconocimiento de signos respiratorios

Una vez determinado el patrón característico del signo respiratorio se procede a buscar su presencia dentro del sonido de audio de la realización completa. Para ello es necesario hacer el acondicionamiento previo a la señal anteriormente descripto. Luego se segmenta en bloques y sobre cada uno se aplica el algoritmo MUSIC para determinar el pseudoespectro correspondiente a cada uno. Finalmente, se genera una *señal de detección* D consistente en la comparación en cada segmento del pseudoespectro obtenido respecto al pseudoespectro patrón del signo patológico. La señal generada es binaria, indicando la presencia (1's) o ausencia de signo detectado (0's) según si la distancia euclídea d_j para el j-ésimo segmento es menor o mayor que un umbral de referencia, respectivamente, según:

$$d_j = \sqrt{\sum_{i=1}^{N}(p_i - q_i)^2}, \tag{1}$$

donde N es el número de muestras del segmento, p el pseudoespectro patrón del signo respiratorio y q el pseudoespectro calculado sobre el segmento.

En la tarea de reconocimiento del signo respiratorio se permitieron diferencias menores a un umbral de máxima diferencia admisible, de modo de discernir entre artefactos en la señal (ruido de autos, etc.). Este umbral se ajusta experimentalmente.

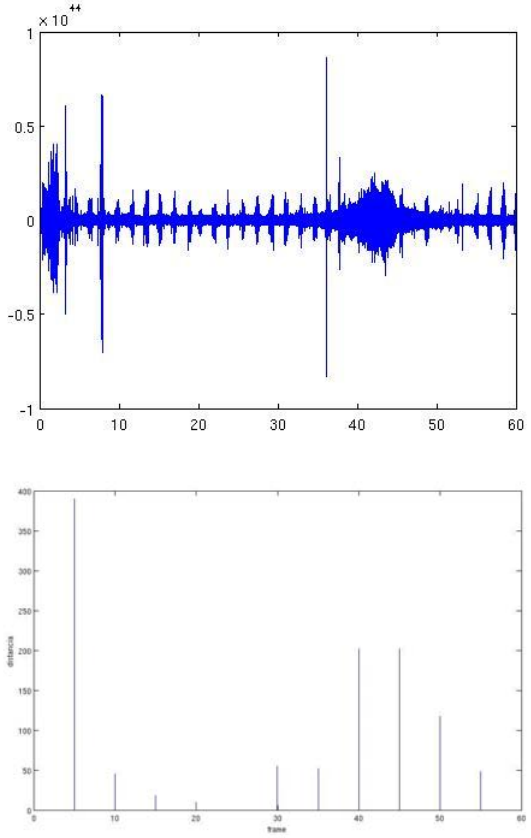


Figura 6. Análisis de distancia entre pseudoespectros. Sonograma de la señal analizada (arriba); distancias calculadas indicadas en el centro de cada frame considerado, sin umbral de selección (abajo).

3. Experimentos y resultados

A efectos de poder evaluar el desempeño del sistema de reconocimiento propuesto, se lo sometió a diferentes situaciones, de modo de poder observar y comparar aspectos en su funcionamiento.

Se observará la respuesta frente a la variación en el número de pollos analizados, provenientes de un lote productivo de 15.000 pollos de 25 días de edad que evidenciaba casos de individuos con signos tempranos de afección respiratoria[3].

[3] Granja avícola localizada en la provincia de Entre Ríos.

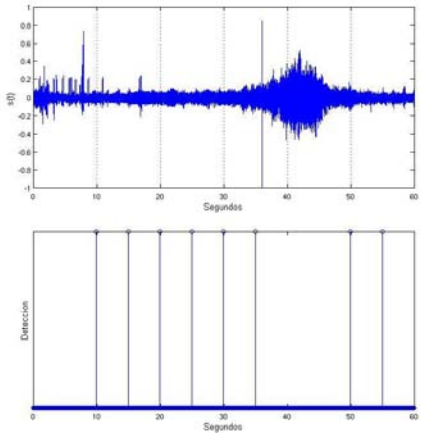


Figura 7. Sonograma de la señal sin procesar de 1 pollo enfermo (arriba) y señal de detección de signos respiratorios (abajo).

3.1. Pruebas con un solo pollo enfermo

En esta instancia se aisló a un pollo con signos respiratorios patológicos, en un recinto alejado del galpón donde se aloja el lote productivo. La distancia del micrófono al pollo fue de alrededor de 10 cm.

La Figura 7 muestra un ejemplo de los resultados obtenidos. Se puede observar cómo el sistema reconoce la presencia del signo respiratorio solamente en aquellos intervalos donde no se presenta ruido, ignorando en este caso la presencia de ruidos al inicio (aproximadamente los primeros 5 s.) y al final la perturbación por un ruido de automóvil (aproximadamente a los 40 s.).

3.2. Pruebas con varios pollos enfermos

En la Figura 8 se puede observar la señal adquirida de un grupo de 4 pollos enfermos y cómo el sistema reconoce el signo respiratorio en aquellos intervalos sin presencia de ruidos anormales.

A diferencia del caso testigo anterior, se observa aquí una mayor periodicidad en los eventos respiratorios de la señal, así como también un incremento en su amplitud. Esto se debe a una respiración parcialmente sincronizada por pequeños subgrupos de pollos, una característica particular de la enfermedad.

3.3. Prueba con mezclas de pollos enfermos y sanos

En la Figura 9 se puede observar el caso de una señal perteneciente a una multitud de 7 pollos enfermos, mezclados con individuos sanos (aproximadamente 10 aves/m^2), registrada con un micrófono colgando 10 cm. sobre los pollos.

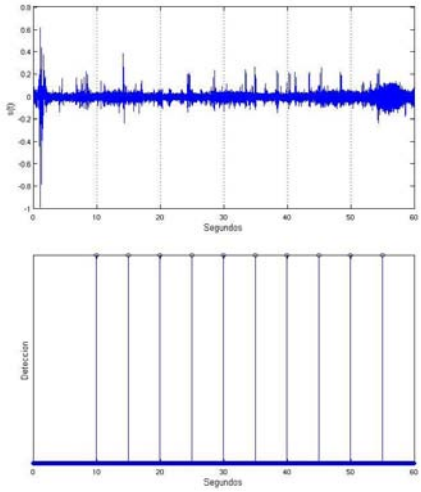


Figura 8. Sonograma de la señal sin procesar de 4 pollos enfermos (arriba) y señal de detección de signos respiratorios (abajo).

En este caso, las condiciones desfavorables de ruido en el entorno dificultan la tarea de detección y fue necesaria una modificación en el umbral empleado anteriormente, reduciéndose el intervalo en el cual el sistema detecta la presencia de signos respiratorios.

4. Conclusiones

En este trabajo se ha presentado el diseño y desarrollo de una técnica computacional de procesamiento de señales de audio que demostró ser de utilidad para la producción avícola, brindando una herramienta automática para la detección temprana de signos respiratorios patológicos.

A partir del registro acústico de los pollos en galpón, empleando técnicas de filtrado y estimación frecuencial, se logró identificar la morfología de signos respiratorios patológicos e identificar la presencia de los mismos en individuos del ambiente productivo.

Una línea de continuación de este trabajo, posterior a la detección del signo patológico, lo constituye la cuantificación estadística de la incidencia de la patología en la población. Este análisis serviría para determinar, mediante un muestreo de la población de un galpón, si la misma presenta o no signos y establecer diferentes niveles de afectación. Por otro lado, en este trabajo se presenta la técnica y experimentación preliminar para demostrar su fiabilidad. Es necesario, por lo tanto, ampliar la experimentación a poblaciones mayores dentro de los galpones, realizando los ajustes necesarios en el sistema para aumentar la robustez en el ambiente natural de producción.

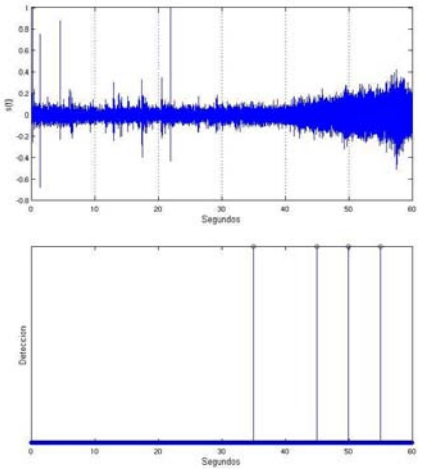


Figura 9. Sonograma de la señal sin procesar de mezcla de pollos enfermos y sanos (arriba) y señal de detección de signos respiratorios (abajo).

Agradecimientos

Los autores desean agradecer a la *Agencia Nacional de Promoción Científica y Tecnológica* (bajo proyecto PAE 37122), la *Universidad Nacional de Litoral* (PACT #58, CAI+D 2011 #58-511, #58-525).

Referencias

1. SENASA. *Plan nacional de sanidad avícola*, 2003.
2. L. Schrader and K. Hammerschmidt. Computer-aided analysis of acoustic parameters in animal vocalisations: a multi-parametric approach. *Bioacoustics*, 7(4):247–265, 1997.
3. Jagmeet S Kanwal, Sumiko Matsumura, Kevin Ohlemiller, and Nobuo Suga. Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *The Journal of the Acoustical Society of America*, 96:1229, 1994.
4. Christopher W Clark. The acoustic repertoire of the southern right whale, a quantitative analysis. *Animal Behaviour*, 30(4):1060–1071, 1982.
5. D. H. Milone, J. Galli, C. E. Martínez, H. L. Rufiner, E. Laca, and C. Cangiano. Reconocimiento automático de sonidos masticatorios en rumiantes. In *Anales de las 37 Jornadas Argentinas de Informática, JII-Agroinformática*, pages 372–384, Santa Fe, Argentina, september 8-12 2008.
6. D. H. Milone, J. Galli, C. Cangiano, H. L. Rufiner, and E. Laca. Automatic recognition of ingestive sounds of cattle based on hidden markov models. *Computers and Electronics in Agriculture*, 87:51–55, sep 2012.
7. Ralph Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34(3):276–280, 1986.
8. L. Marple. *Digital Spectral Analysis With Applications*. Prentice-Hall, 1987.

Sievert-type measurement and acquisition system for the study of hydrogen storage in solids

Jorge Runco, Marcos Meyer
IFLP – Departamento de Física – Facultad de Ciencias Exactas – UNLP
CONICET
{ runco, meyer }@fisica .unlp.edu.ar

Abstract
This paper shows the development and implementation of a system to determine and analyze parameters of interest in the study of hydrogen absorption and desorption mechanisms in solids using the Sievert volumetric method. The experiment is controlled through a PC-type computing system by automatically measuring, controlling, recording and graphing the evolution of variables (pressures and temperatures) according to a set of previously programmed parameters. The manual monitoring, adjustment and operation option is also available to tune the experiment. The software was developed in a high-level programming language (Delphi) which offers the user a graphical interface typical of visual languages. In addition, results for applying the present system to typical ternary hydrides are presented.

Key words: hydrogen storage, processes, absorption and desorption, data acquisition.

1. Introduction

One of the most important challenges for the development of hydrogen utilization as an energy vector is the possibility of storing it in a safe and effective way [1].
Hydrogen in a gaseous state occupies a very large volume and requires very high pressures in storage reservoirs whereas, in a liquid state, it needs reservoirs at very low temperatures.
Since hydrogen is highly reactive, there is a significant number of elements capable of reacting with it to form hydrides under appropriate pressure and temperature conditions. Hence, hydrogen storage in a solid state appears as a more effective alternative in terms of volume with respect to the other methods mentioned above.
Absorption in metals, which forms a hydride phase, has many advantages over current systems (compression and liquefaction), since it does not require compressing and liquefying tasks or cryogenic temperatures.
An important aspect in experimental research is the analysis of hydrogen absorption-desorption properties of new materials, as well as the study of absorption and desorption kinetics.
As considerable pressure changes involve relatively small mass quantities, using volumetric techniques is especially suitable for this kind of research (mainly considering that hydrogen is the lightest element of all).

2. Measurement system description

The equipment is based on a Sievert-type system [2], which allows studying hydrogen absorption-desorption kinetics at different temperatures, keeping pressure level constant in the reaction chamber, in a wide range of temperatures (from 300 K to 1000 K) and pressures (from 1 mbar to 50 bar).

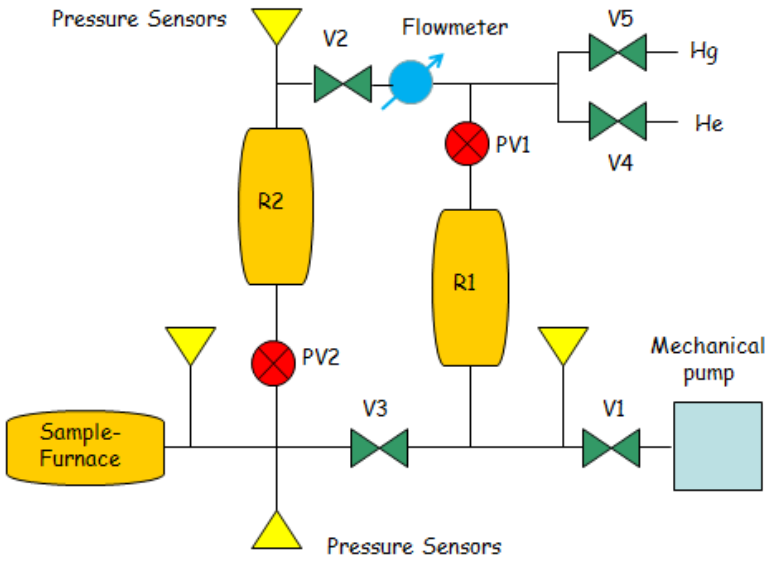


Figure 1. Schematic diagram of the Sievert apparatus.

The measuring instrument was developed based on a PC-type computing system which incorporates a 12-bit general purpose A/D interface [3], 5 inputs of which are used to measure analog magnitudes –4 for pressures and 1 for temperature. It also has digital outputs for the automatic opening and closing of the 7 solenoid valves (V1, V2, V3, V4, V5, PV1 and PV2) of the measurement system. The schematic diagram above illustrates the location of pressure sensors [4] in the experiment as well as valves automatically actuated by the control system.

The measurement system has a furnace and a temperature controller to conduct the experiment at different sample temperatures. This controller communicates with the computer by means of an RS-232 interface and MODBUS protocol [5].

A flowmeter was added to the system to control and measure gas flow during the experiment. This device communicates with the computer also through an RS-232 interface and the MODBUS protocol.

The software developed makes it possible to open and close solenoid valves at any desired time, to acquire and store parameters of interest –pressures, temperature and flow– and to preset temperature and flow values (setpoints) before starting the experiment. In addition, routines necessary to communicate with furnace and gas flow controllers through the MODBUS protocol were implemented.

During the course of the experiment, the abovementioned parameters are acquired and stored, valves open and close when pressure conditions set before starting the measurement are met or for controlling gas flow with the flowmeter, and the system is capable of performing the experiment at a constant temperature, with linearly increasing temperature (temperature ramp), when cooling, or with a combination of these stages.

Figure 2 shows the interconnection between the different modules and components. The Signal Adaptation Electronics module is in charge of conditioning (instrumentation amplifiers) signals coming from sensors at the input of the A/D converter. Moreover, this module is equipped with the circuits needed to actuate the valves with digital signals since they operate with 220 V. Isolation between this voltage and the computing system is required; this is accomplished with optically-coupled circuits.

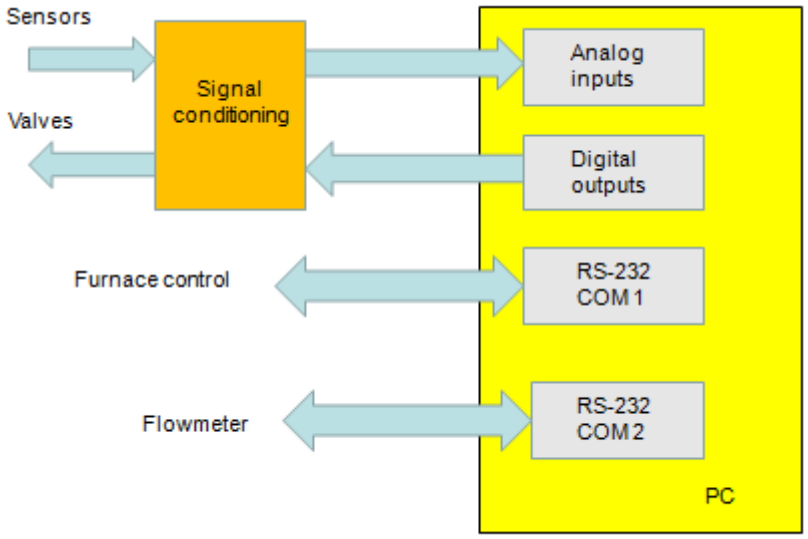


Figure 2. Acquisition and control diagram.

3. Results

The equipment described above was used in several experiments for studying the kinetics of absorption and desorption to form typical ternary hydrides.

Figure 3 displays temperature and evolution of the material's pressure when releasing hydrogen according to time. When preset pressure conditions are met, valves automatically open and close between these two values. When hydrogen is released, pressure increases up to a certain value, hydrogen is discharged, pressure decreases, and the cycle is repeated until the sample releases all the hydrogen. Figure 4 illustrates the analytical treatment of measured data.

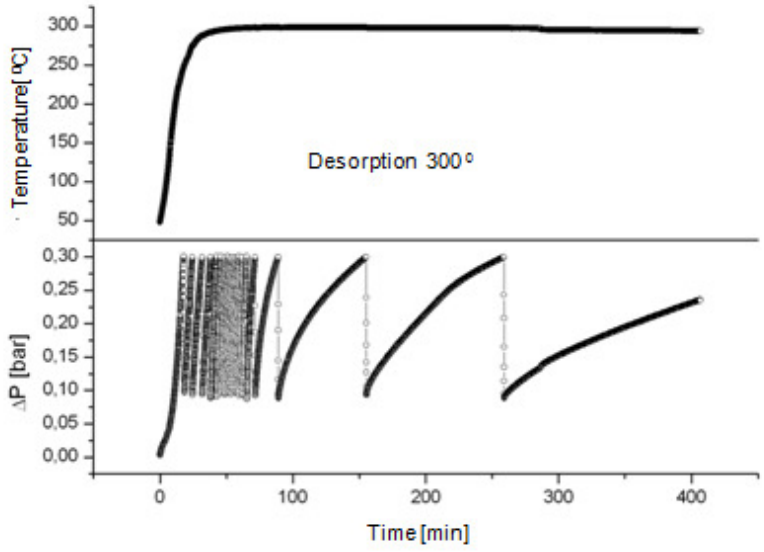


Figure 3. Temperature according to time. Sample desorption. Pressure variation according to time.

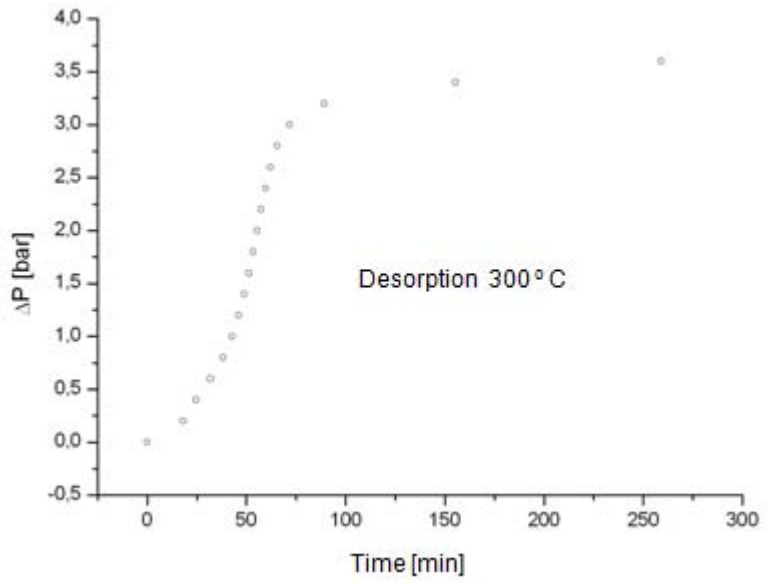


Figure 4. Desorption data treatment.

4. Conclusions

This paper describes the development and implementation of a piece of equipment for studying hydrogen absorption and desorption in metals. Having used commercially available components, the importance of this proposal lies in its low cost.

The design complies with the requirements specified by a research group from the IFLP (Instituto de Física de La Plata [***La Plata Physics Institute***] - CONICET [***for its Spanish acronym, National Council of Scientific and Tehnical Research***), Departamento de Física, Facultad de Ciencias Exactas, UNLP [***Physics Department, Faculty of Exact Sciences, National University of La Plata***]) that is part of the project "Materiales nanoestructurados de aplicación en energías alternativas: síntesis, caracterización y modelado" [***"Nanostructured materials applicable to alternative energies: synthesis, characterization and modeling"***].

The design currently in operation went through several stages: from the development and implementation of the signal adaptation module, to the software that controls and automates the experiment, up to the present situation with variations from the original experiments: constant temperature test, with (linear) increase, acquisition during cooling and gas flow measurement and control.

5. References

[1] Optimización de un hidruro complejo para almacenamiento de hidrógeno. Ph.D. thesis. June, 2009 – Cardozo, César Luis - Centro Atómico Bariloche [***Bariloche Atomic Center***].

[2] R. Checchetto, G. Trettel and A. Miotello. 2003 – Sievert-type apparatus for the study of hydrogen storage in solids. Measurement Science and Technology.

[3] Conversor A/D- ADQ-12 – Microaxial

[4] Druck-PDCR 4000 Series – High Performance Millivolt Output Pressure Transducers. Motorola - MPX2200 Series – On-Chip Temperature Compensated & Calibrated Pressure Sensors.

[5] MODBUS Protocol Specification - http://www.modbus.org.

II WORKSHOP DE SEGURIDAD INFORMÁTICA

- WSI -

CACIC2013

II WORKSHOP DE SEGURIDAD INFORMÁTICA

- WSI -

ID	Trabajo	Autores
5606	**Improving versatility in keystroke dynamic systems**	Enrique P. Calot (UBA), Juan Manuel Rodríguez (UBA), Jorge Salvador Ierache
5713	**Gestión Integral de Seguridad de Infraestructuras críticas para las organizaciones locales alineadas a las Normas IRAM ISO IEC 27.001 y 27002.**	Mirta Elizabeth Navarro (UNSJ), María del Carmen Becerra (UNSJ)
5753	**Model Design for a Reduced Variant of a Trivium Type Stream Cipher**	Antonio Castro Lechtaler (IESE), Marcelo Cipriano (IESE), Edith García (IESE), Julio Liporace (IESE), Ariel Maiorano (IESE), Eduardo Malvacio (IESE)
5801	**Usability Support Security Patterns**	Susana Romaniz (UTN-FRSF), Marta Castellaro (UTN-FRSF), Juan Ramos (UTNFRSF), Ignacio Ramos (UTN-FRSF)
5873	**Analizador de Intents en Android**	Joaquín Erario (UNRC), Christian Rovera (UNRC), Francisco Bavera (UNRC)

Improving versatility in keystroke dynamic systems

Enrique Calot, Juan Manuel Rodriguez, and Jorge Salvador Ierache*

Laboratorio de Sistemas de Información Avanzados,
Facultad de Ingeniería, Universidad de Buenos Aires,
Buenos Aires, Argentina
{ecalot,jmrodri}@fi.uba.ar,jierache@yahoo.com.ar

Abstract. Keystroke dynamics is a biometric technique to identify users based on analyzing habitual rhythm patterns in the way they type. In order to implement this technique different algorithms to differentiate an impostor from an authorized user were suggested. One of the most precise method is the Mahalanobis distance which requires to calculate the covariance matrix with all that this implies: time processing and track each individual keystroke event. The hypothesis of this research was to find an algorithm as good as Mahalanobis which does not require track every single keystroke event and improve, where possible, the processing time. To make an experimental comparison between Mahalanobis distance and euclidean normalized, a distance which only requires calculate the variance, an already studied dataset was used. The results were that use normalized euclidean distance is almost as good as Mahalanobis distance even in some cases could work better.

Keywords: Keystroke Dynamics, Web based authentication, Mahalanobis distance, Biometrics, typing biometrics

1 Introduction

The variables that help make a handwritten signature a unique human identifier also provide a unique digital signature in the form of a stream of latency periods between keystrokes. The handwritten signature has a parallel on the keyboard. The same neurophysiological factors that make a written signature unique are also exhibited in a user's typing pattern[1].

Password typing is the most widely used identity verification method in World Wide Web based electronic commerce. Due to its simplicity, however, it is vulnerable to impostor attacks. Keystroke dynamics and password checking can be combined to result in a more secure verification system[2].

This authentication is fragile when there is a careless user and/or a weak password. Biometric characteristics are unique to each person and have advantages as they could not be stolen, lost, or forgotten[3, 4].

* This paper was done with Cloodie R&D Support

The biometric technology employed in this paper is the typing biometrics, also known as keystroke dynamics. Typing biometrics is a process that analyzes the way a user types at a terminal by monitoring the keyboard inputs in attempt to identify users based on their habitual typing rhythm patterns[5, 4].

Even though WWW keystroke dynamic systems may run locally on the web browser, due to security measures it should be ran on the webserver layer. This paper discusses an approach to reduce data transmission size.

Using a know dataset[6] we designed an experiment to compare three methods to compute the keystroke dynamics of users and compare them with impostors.

Our hypothesis is that one of the most used and precise method –the Mahalanobis distance– is as successful as the second method –normalized euclidean distance–. Ignoring the success rates, there are some advantages that the normalized euclidean distance has over the Mahalanobis distance, so if the hypothesis is confirmed using this method should prove to be a more useful way to calculate keystroke dynamics.

Some advantages of the normalized euclidean distance ar the lesser transferred information, processing time and the bigger versatility when changing passwords.

2 Current implementations

There are different methods to compare keystrokes, all based on measuring the distances between two strokes, a negative result is found when both differ more than a threshold. One of the best methods is the Mahalanobis distance[2, 6].

Three methods are shown below, each method is a generalization of the previous one.

2.1 Euclidean distance

The time a user press a key or the time between one key and the other may result in a vector of events ($\boldsymbol{\Gamma}$). Each event represent a key hold time or the elapsed time between two keys. Since in the training phase an event may occur several times, the vector is a list of the expected values of every event time.

Calculating the euclidean distance between two vectors works as a comparison algorithm, with relatively high success rates[6].

$$d(\boldsymbol{\Gamma_1}, \boldsymbol{\Gamma_2})^2 = \|\boldsymbol{\Gamma_1} - \boldsymbol{\Gamma_2}\|^2 = \sum_{i=1}^{N} (\Gamma_{1,i} - \Gamma_{2,i})^2 \tag{1}$$

Where $\boldsymbol{\Gamma_1}$ is the vector of training event times and $\boldsymbol{\Gamma_2}$ is the vector of testing event times.

To optimize calculation timings the squared norm is actually used.

2.2 Normalized euclidean distance

A disadvantage of the former method is that important information is ignored. The variance of each event time should be taken into consideration, and that is exactly what the normalized euclidean distance does: adding the variance (s_i^2) of each event time.

Using the inverted variance of the training set ($\boldsymbol{\Gamma_1}$) as a weight factor, the normalized euclidean distance is defined as

$$d(\boldsymbol{\Gamma_1}, \boldsymbol{\Gamma_2})^2 = \sum_{i=1}^{N} \frac{(\Gamma_{1,i} - \Gamma_{2,i})^2}{s_i^2} \tag{2}$$

where s_i^2 is the variance of each element of $\boldsymbol{\Gamma_1}$.

2.3 Mahalanobis distance

Again, the former method is skipping information, this time is the covariance between events.

Mahalanobis distance is defined as

$$d(\boldsymbol{\Gamma_1}, \boldsymbol{\Gamma_2})^2 = (\boldsymbol{\Gamma_1} - \boldsymbol{\Gamma_2})^T S^{-1} (\boldsymbol{\Gamma_1} - \boldsymbol{\Gamma_2}) \tag{3}$$

Where S^{-1} is the inverted covariance matrix corresponding to all events in the training set $\boldsymbol{\Gamma_1}$[7].

3 Problems of Mahalanobis distance in keystroke dynamics

Translating each method to a kernel matrix it turns out that in equation 3 the matrix S is the identity in the euclidean distance, a diagonal with the variances in the normalized euclidean distance and the covariance matrix in the Mahalanobis distance.

3.1 Distance kernel matrix size

To generate the covariance matrix for the Mahalanobis distance all key-press events and their respective timings should be transmitted to the server while training –or at least the covariance matrix and the expected event timings–. But to generate the diagonal matrix for the normalized euclidean method is it possible to send only three integer numbers per event or two floats.

Using the property $Var[X] = E[X^2] - E[X]^2$ it is possible to generate the variance using only the sum of squares $SS = \sum_{i=0}^{n} \Gamma_{1,i}^2$, the sum $S = \sum_{i=0}^{n} \Gamma_{1,i}$ and the total n since $E[X^2] = \frac{SS}{n}$ and $E[X] = \frac{S}{n}$. All three numbers are natural and may be combined in an $\mathbb{N}^3$ vector which supports commutative addition properties. This method allows parallelized variance calculus[9].

Table 1. Different parameters to be sent to the server

Distance Method	Variables
Euclidean	$(S, n) \in \mathbb{N}^{2\times n}$ or $\Gamma = E[X] \in \mathbb{R}^n$
Normalized euclidean	$(S, SS, n) \in \mathbb{N}^{3\times n}$
Mahalanobis	$\Gamma = E[X] \in \mathbb{R}^n$ and $Cov[X] \in \mathbb{R}^{n\times n}$

There are several ways to send the data depending on the algorithm to be used. Table 1 compares them.

For example, when 20 events are used, the covariance matrix has $20\times20 = 400$ values and the $\Gamma = E[X]$ vector has 20 values. Normally a $\mathbb{R}^{n\times n}$ matrix can be encoded with n^2 numbers, but as $Cov[a, b] = Cov[b, a]$, covariance matrix is symmetric and therefore it can be encoded with $\frac{n(n+1)}{2}$ numbers. Assuming a real number and an integer has the same size, the transmission would be of $\frac{20\times21}{2} + 20 = 230$ numbers while the normalized euclidean only transmits $3 \times 20 = 60$ numbers. Generalizing, the data transmission of Mahalanobis distance is $\frac{n(n+1)}{2} + n$ reals, that is $\mathcal{O}(n^2)$, normalized euclidean is $3n$ integers, that is $\mathcal{O}(n)$ and euclidean is $2n$ integers or n reals, that is also $\mathcal{O}(n)$.

3.2 One password algorithm

Another problem is that training is done with only one password. A new password should require the user to re-train all the covariance matrix with Mahalanobis.

Normalized euclidean distance may reuse the variances of the common keys between two different passwords while Mahalanobis distance may not.

3.3 Backspace eliminating digraphs

When the user trains it is possible that mistakes a character and use backspace to correct it, in this case one event will be missing. For example the word "train" has 5 characters so the events will be `t.hold`, `t.up-r.down`, `r.hold`, etc. The problem occurs when "te`[backspace]`rain" is typed, the event `t.up-r.down` was not recorded because there were two keys in the middle "e" and `[backspace]`.

Having a variable number of events per key is a problem to calculate the covariance matrix, but allows to process backspaces in passwords (sacrificing the success rate due to lesser information available) and free text.

Table 2 shows an example of Mahalanobis method with three pairs of events and normalized euclidean with three and two times per event respectively.

3.4 Processing times

Calculating the covariance matrix and inverting it should take a considerable amount of time for the Mahalanobis method, the time here is expected to be

Table 2. Example of how timing counts are dependent on the event in Mahalanobis distance

Method	Key	Times	Matrix S	Inverse S^{-1}
Mahalanobis	`A.hold` `A.up-B.down`	$\begin{Bmatrix} 90 \\ 161 \end{Bmatrix}, \begin{Bmatrix} 99 \\ 171 \end{Bmatrix}, \begin{Bmatrix} 97 \\ 174 \end{Bmatrix}$	$\begin{bmatrix} \frac{67}{3} & \frac{175}{6} \\ \frac{175}{6} & \frac{139}{3} \end{bmatrix}$	$\begin{bmatrix} \frac{556}{2209} & -\frac{350}{2209} \\ -\frac{350}{2209} & \frac{268}{2209} \end{bmatrix}$
Normalized euclidean	`A.hold` `A.up-B.down`	{90}, {99}, {97} {161}, {171}	$\begin{bmatrix} \frac{67}{3} & 0 \\ 0 & 50 \end{bmatrix}$	$\begin{bmatrix} \frac{3}{67} & 0 \\ 0 & \frac{1}{50} \end{bmatrix}$

$\mathcal{O}(n^2)$. Normalized euclidean should also take time to compute the variances, but this procedure is $\mathcal{O}(n)$. Inverting the matrix lacks of relevant costs due to the properties of the diagonal matrices. Euclidean distance should be the fastest algorithm because of its simplicity.

It is important to remark that due to parallelized calculation of the variance, part of the calculating time in training mode for the normalized euclidean distance may be done while reading the keyboard by the user machine.

The experiment also intends to measure algorithms processing time.

4 Experimental comparison

We use an already studied dataset for two main reasons, one is because it was collected in a controlled laboratory environment, the second reason is because 14 detection algorithms were tested using this dataset[6] and that give us a big framework to start our research. The data was collected from 50 different users along 8 days or sessions –totalizing 400 cases per user–, in each session the users typed always the same string: "`.tie5Roanl`" which represents a reasonable secure password. When any error in the sequence was detected, the subject had been prompted to retype the password. To make this a laptop was set up with an external keyboard to collect data and a Windows application was developed to prompt the subjects to type the password. The application displays it in a screen with a text-entry field. In order to advance to the next screen, the subject must type the 10 characters of the password correctly in sequence and then press Enter. The data set contains the hold time of each key, the time between two consecutive keys were pressed and the time since one key was released and the next was pressed. One of the three values depends linearly of the other two. Due to preconditions of covariance one value was dropped away leaving two values per key.

From the 400 cases per user, the first 200 cases were used to train the detection algorithm and the second 200 cases were used to validate it, also the first 5 cases of all the other users were taken to generate an impostor dataset in order to validate negative cases. This data set and schema was taken from Killourhy and Maxion[6].

We performed 19 tests, the first using two events (the first two values of the Γ vector) and increasing the number of events until the last one, using all twenty.

We expected to have a best success rate in the last test because it counts with more information. We ran the three mentioned algorithms in each test.

Finally we calculated the area under the receiver operating characteristics (ROC) curve –a performance measure for machine learning algorithms commonly used in systems that learns by being shown labeled examples[8]–. This method, known as AUC, was chosen because it is a classifier performance evaluator independent of the decision threshold chosen on the keystroke distances.

5 Results

With the one key test case we obtained in one sample user $\boldsymbol{\Gamma} = [98.98, 166.905]$, where first value corresponds to the expected key-hold time and the second to the expected elapsed time until the next key was pressed. Both times are expressed in milliseconds.

$$S^{-1}_{Mahalanobis} = Cov[\boldsymbol{\Gamma}]^{-1} = \begin{bmatrix} 341.29 & 282.19 \\ 282.19 & 5464.9 \end{bmatrix}^{-1} = \begin{bmatrix} 0.0031 & -0.00016 \\ -0.00016 & 0.0002 \end{bmatrix}$$

$$S^{-1}_{normalizedEuclidean} = \begin{bmatrix} 341.29 & 0 \\ 0 & 5464.9 \end{bmatrix}^{-1} = \begin{bmatrix} 0.0029 & 0 \\ 0 & 0.00018 \end{bmatrix}$$

$$S^{-1}_{euclidean} = S_{euclidean} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Note that $S_{Mahalanobis}$ and $S_{normalizedEuclidean}$ have the same diagonal values but this is not the case with their inverses.

Table 3. Experimental results

N	Method	Total Errors	ROC	Zero-miss	False-Alarm	Time
2	Mahalanobis	0.01887	80.43%	7461 of 12750	769 of 10200	1.356s
2	Normalized euclidean	0.01899	80.17%	7451 of 12750	788 of 10200	1.300s
2	Euclidean	0.02240	70.61%	9030 of 12750	649 of 10200	0.872s
20	Mahalanobis	0.00970	94.60%	5576 of 12750	464 of 10200	1.896s
20	Normalized euclidean	0.00919	94.84%	5581 of 12750	428 of 10200	1.764s
20	Euclidean	0.01440	88.27%	6853 of 12750	844 of 10200	1.704s

Table 3 shows the results of the 3 methods with 2 and 20 timing events respectively. Each method shows the area under ROC curve in percentage among with zero-miss and false-alarm rates. It is also shown the total processing time

of training, testing all the 12750 positive and 10200 negative sets and calculating the results.

In the last test –with 20 events–, normalized euclidean distance method performed even better than Mahalanobis.

As expected, our hypothesis that in the test with 20 timing events is better than the test with 2 was confirmed and that Mahalanobis and normalized euclidean distance are both superior than euclidean distance was confirmed too. Processing is, as expected, bigger for Mahalanobis and decreasing for normalized euclidean and finally, the fastest method, euclidean distance.

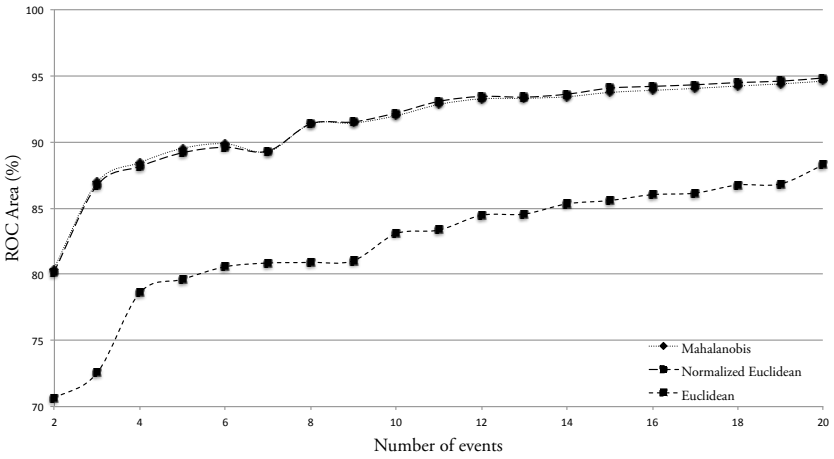


Fig. 1. Distance methods compared in success versus amount of information

In Figure 1 it is possible to appreciate how similar are the normalized euclidean and Mahalanobis distances compared to the euclidean.

6 Conclusions

Normalized euclidean distance and Mahalanobis distance are almost the same in all ran tests. In the case of 20 events the results varied 0.24%. Normalized euclidean was faster than Mahalanobis distance for $132ms$ but slower than euclidean for only $60ms$. Versatility in normalized euclidean is also an advantage, passwords may be changed and the already-trained keys be kept in the new training. Those results lead to the conclusion that normalized euclidean distance is strong enough to be used and its advantages in data sizes and versatility are considerably important to be chosen against Mahalanobis distance and its success rate suggests that it should be employed against euclidean distance.

6.1 Future lines of research

We are exploring the way users may vary keystroke dynamics over the time. Using variance parallelization principle[9] there is a way to "forget" the training, making it autoadaptive with this time-wise learning technique. We are also exploring new fields on keystroke dynamics that include user emotional state detection.

Acknowledgments. This paper acknowledges support from Cloodie R&D.

References

1. Joyce, R.; Gupta, G.: Identity authentication based on keystroke latencies. Commun. ACM 33, 2 (February 1990), 168-176. http://doi.acm.org/10.1145/75577.75582 (1990)
2. Cho, S.; Han, C.; Han., D. H.; Kim, H. I.: Web based Keystroke Dynamics Identity Verification using Neural Network. Journal of Organizational Computing and Electronic Commerce, Vol. 10, No. 4, pp. 295–307 (2000)
3. Polemi, D.: Biometric Techniques: Review and Evaluation of Biometric Techniques for Identification and Authentication, Including an Appraisal of the Areas Where They are Most Applicable. Institute of Communication and Computer Systems, National Technical University of Athens, Athens, Greece. Retrieved on 2013-07-01: ftp://ftp.cordis.lu/pub/infosec/docs/biomet.doc, EU Commission Final Rep. (1997)
4. Araujo, L. C. F.; Sucupira, L. H. R.; Lizarraga, M. G.; Ling, L. L.; Yabu-Uti, J. B. T.: User authentication through typing biometrics features. Signal Processing, IEEE Transactions on, vol. 53, no. 2, pp.851–855 (2005)
5. Monrose, F.; Rubin, A. D.: Keystroke dynamics as a biometric for authentication. Future Gen. Comput. Syst., vol. 16, no. 4, pp. 351-359 (2000)
6. Killourhy, K. S.; Maxion, R. A.: Comparing Anomaly-Detection Algorithms for Keystroke Dynamics. In International Conference on Dependable Systems & Networks (DSN-09), pp. 125–134, Estoril, Lisbon, Portugal, 29 June to 02 July 2009. IEEE Computer Society Press, Los Alamitos, California (2009)
7. Mahalanobis, P. C.: On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2 (1): 49-55. Retrieved on 2013-07-01: http://www.new.dli.ernet.in/rawdataupload/upload/insa/INSA_1/20006193_49.pdf (1936)
8. Bradley, A. P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, Volume 30, Issue 7, July 1997, Pages 1145–1159, ISSN 0031-3203, http://dx.doi.org/10.1016/S0031-3203(96)00142-2. (1997)
9. Chan, T. F.; Golub, G. H.; LeVeque, R. J.: Updating Formulae and a Pairwise Algorithm for Computing Sample Variances. Technical Report STAN-CS-79-773, Department of Computer Science, Stanford University. Retrieved on 2013-07-01: ftp://reports.stanford.edu/pub/cstr/reports/cs/tr/79/773/CS-TR-79-773.pdf (1979)

Gestión Integral de Seguridad de Infraestructuras críticas paralas organizaciones locales alineadas a las Normas IRAM ISO IEC 27.001 y 27002.

Mag. Licenciada Mirta Elizabeth Navarro[1] Mag. Abogado María del C. Becerra[2],

marisabecerra2005@yahoo.com.ar, mirthaenavarro@yahoo.com.ar

Abstract.: Con esta propuesta de gestión integral diseñada para la protección de la información alineada a la norma IRAM ISO IEC 27001, 27002 y a lastendencias y normas de seguridad informática, consistente en la creación y adopción de un marco regulatorio que favorezca la identificación y protección de las infraestructuras estratégicas y críticas de la U.N.S.J., se favorecerá la colaboración entre los diversos sectores públicos y privados, para el desarrollo de estrategias y estructuras adecuadas para la protección de los activos de información. Todo ellodará real importancia a la nueva visión del estado, liderada por la incorporación de las TIC´s en los procesos administrativos y bajo el sustento de la comunicación íntegra y confidencial necesaria en el entorno globalizado de e-gobierno, se busca dar impulso a la administración electrónica.

Keywords: Gestión Integral de seguridad. Infraestructuras críticas. Propuesta para implementar el plan Infraestructuras críticas y ciberseguridad.

1 Introducción

Para ayudar a garantizar una gestión de integral de las infraestructuras críticas[3] de las empresas se tomaron en cuenta dos normas de la familia de las normas ISO 27000 adaptadas y traducidas en nuestro país, especialmente las Normas IRAM ISO 27.001[4] y 27002[5] y su antecedente 17799), en base a ellas se definieron los requisitos para el sistema de gestión de seguridad (SGSI) propuesto.

En el proyecto "Convergencia de Tecnologías informáticas y Metodologías para la implementación de sistemas de Información"se analizaron las políticas, prácticas, procedimientos y estructuras organizacionales como conjunto de controles necesarios para implementar un sistema de gestión para las infraestructuras críticas.

En Argentina por Resolución 580/2011 se crea, en el ámbito de la Oficina Nacional de Tecnologías de Información de la subsecretaria de Tecnologías de Gestión de la Secretaria de Gabinete de la Jefatura de Gabinete de Ministros, el "Programa Nacional De Infraestructuras Criticas De Información y Ciberseguridad" en el marco de lo establecido la Ley de Ministerios (t.o. Decreto Nº 438/92), a fin de impulsar la creación y adopción de un marco regulatorio específico que propicie la identificación y protección de las infraestructuras

[1]Licenciada en Administración de empresas egresada de la U.N.S.J. Magíster en Gestión de Organizaciones egresada de la Universidad de Valparaíso. Chile. Docente de la U.N.S.J. Directora del Proyecto Convergencia de Tecnologías informáticas y Metodologías para la implementación de Sistemas de información

[2]Abogado, egresado de la UCC. Magíster en Informática egresado de la Universidad Nacional de la Matanza. Docente Investigadora de la U.N.S.J. Directora del Instituto de informática del Foro de Abogados de San Juan

[3]Las infraestructuras críticas son aquellas instalaciones, redes, equipos físicos y de tecnología de la información cuya interrupción o destrucción tendría un impacto en el bienestar de los ciudadanos o en el eficaz funcionamiento del gobierno. Las infraestructuras críticas están presentes en numerosos sectores: financiero, transporte y distribución, energía, salud, comunicaciones, y administraciones públicas.

[4] En Argentina es IRAM, como organismo nacional de normalización, quien la estudia a través del Subcomité de Seguridad de la Información y la adopta como IRAM-ISO/IEC 27001. Se publica bajo el nombre Tecnología de la información. Sistemas de gestión de la seguridad de la información (SGSI). Requisitos, difundiéndola en la región a través de cursos y seminarios.

[5]Aprobada y consensuada por el IRAM (Instituto de Normalización Argentino) en el año 2002

estratégicas y críticas del Sector Público Nacional, los organismos interjurisdiccionales y las organizaciones civiles y del sector privado que así lo requieran, y la colaboración de los mencionados sectores con miras al desarrollo de estrategias y estructuras adecuadas para un accionar coordinado hacia la implementación de las pertinentes tecnologías, entre otras acciones.

El "Programa Nacional de Infraestructuras Criticas de Información y Ciberseguridad" no interceptará ni intervendrá en conexiones o redes de acceso privado de acuerdo a lo estatuido por la Ley Nº 25.326 de Protección de los Datos Personales y su Decreto Reglamentario Nº 1558 del 29 de noviembre de 2001.

El programa de ICIC, tendrá a su cargo los siguientes objetivos:

a) Elaborar y proponer normas destinadas a incrementar los esfuerzos orientados a elevar los umbrales de seguridad en los recursos y sistemas relacionados con las tecnologías informáticas en el ámbito del Sector Público Nacional.

b) Colaborar con el sector privado para elaborar en conjunto políticas de resguardo de la seguridad digital con actualización constante, fortaleciendo lazos entre los sectores público y privado; haciendo especial hincapié en las infraestructuras críticas.

c) Administrar toda la información sobre reportes de incidentes de seguridad en el Sector Público Nacional que hubieren adherido al Programa y encausar sus posibles soluciones de forma organizada y unificada.

d) Establecer prioridades y planes estratégicos para liderar el abordaje de la ciberseguridad, asegurando la implementación de los últimos avances en tecnología para la protección de las infraestructuras críticas.

e) Investigar nuevas tecnologías y herramientas en materia de seguridad informática.

f) Incorporar tecnología de última generación para minimizar todas las posibles vulnerabilidades de la infraestructura digital del Sector Público Nacional.

g) Asesorar a los organismos sobre herramientas y técnicas de protección y defensa de sus sistemas de información.

h) Alertar a los organismos que se adhieran al presente Programa sobre casos de detección de intentos de vulneración de infraestructuras críticas, sean estos reales o no.

i) Coordinar la implementación de ejercicios de respuesta ante la eventualidad de un intento de vulneración de las infraestructuras críticas del Sector Público Nacional.

j) Asesorar técnicamente ante incidentes de seguridad en sistemas informáticos que reporten los organismos del Sector Público Nacional que hubieren adherido.

k) Centralizar los reportes sobre incidentes de seguridad ocurridos en redes teleinformáticas del Sector Público Nacional que hubieren adherido al Programa y facilitar el intercambio de información para afrontarlos.

I) Actuar como repositorio de toda la información sobre incidentes de seguridad, herramientas, técnicas de protección y defensa.

m) Promover la coordinación entre las unidades de administración de redes informáticas del Sector Público Nacional, para la prevención, detección, manejo y recopilación de información sobre incidentes de seguridad.

n) Elaborar un informe anual de la situación en materia de ciberseguridad, a efectos de su publicación abierta y transparente.

ñ) Monitorear los servicios que el Sector Público Nacional brinda a través de la red de Internet y aquellos que se identifiquen como Infraestructura Crítica para la prevención de posibles fallas de Seguridad.

o) Promover la concientización en relación a los riesgos que acarrea el uso de medios digitales en el Sector Público Nacional, las Organizaciones de Gobierno, al público en general, como así también del rol compartido entre el Sector Público y Privado para el resguardo de la Infraestructura Crítica.

p) Difundir información útil para incrementar los niveles de seguridad de las redes teleinformáticas del Sector Público Nacional.

k) Interactuar con equipos de similar naturaleza.

II.-Gestión integral de la seguridad.

A los efectos de garantizar la selección de controles de seguridad adecuados y proporcionales y para proteger la información crítica de las organizaciones se eligieron las Normas IRAM ISO IEC mencionadas porque se consideran recomendables para cualquier empresa grande o pequeña en cualquier parte del mundo y para aquellos sectores que tienen información crítica o gestionan la información de otras empresas.

Para protegerse de estas amenazas la British Standards Institution publicó en 1995 la norma BS 7799 - parte 1 y 2- que sirvió como antecedente a ISO para el estudio de dos estándares internacionales adoptados y reconocidos a nivel mundial: la norma ISO/IEC 17799:2000 Information technology - Security techniques - Code of practice for information security management, revisada en 2005 y reemplazada por la ISO/IEC 27002, que establece los requisitos fundamentales a tener en cuenta para establecer, operar, controlar, revisar, mantener y mejorar un sistema de gestión de seguridad de la información. La norma 27.002 establece un conjunto de reglas de normalización de los conceptos y operatorias de seguridad informática.

Tal cual lo enfatizan los autores, la seguridad de la información se logra implementando un conjunto adecuado de controles que abarca políticas, practicas, y procedimientos, estructuras organizaciones y funciones del software. Estos controles deben ser establecidos para garantizar que se logren los objetivos específicos de seguridad de la organización".

La edición 2000 de esta norma fue publicada por IRAM dando como resultado la IRAM-ISO/IEC 17799:2002. Ésta fue estudiada por el Subcomité de Seguridad de la Información de IRAM (revisión de la IRAM 17798 cuyo antecedente es la BS 7799-2).

La norma está basada en el mismo modelo de los sistemas de gestión de la calidad de la familia ISO 9000. Establece conceptos similares sobre Requisitos de la documentación: alcance, política de seguridad, enfoque sistémico de identificación y valoración de riesgos, operaciones para el tratamiento de los riesgos, objetivos de control, controles y aplicabilidad de los mismos.

Actualmente para certificar un Sistema de Gestión de la Seguridad implementado bajo la norma ISO/IEC 17799 (ISO 27002) se utiliza la norma ISO/IEC 27001: 2005.

Tal como lo prevé la Norma IRAM ISO IEC 27.001 para una adecuada gestión de la seguridad de la información se propone implantar en las organizaciones locales un sistema que aborde esta tarea de una manera metódica, documentada y basada en objetivos claros de seguridad y en una evaluación de los riesgos a los que está sometida la información de las organizaciones locales.

En la práctica esta norma se presenta embebida en varias normativas estatales en Argentina. Forma parte de reglamentaciones de organismos del Estado para cumplir con diversos procedimientos, entre los que se destaca la comunicación A4609 del BCRA (Banco Central de la República Argentina). Un decreto del Poder Ejecutivo de la Secretaria de la Función Pública (2006) instaló la norma con el nombre de Sistema de Gestión de Seguridad del Estado. Existen 400 organismos estatales que la implementan pero no la certifican.

La comunicación A-4609 del Banco Central de la República Argentina[6], resulta así aplicable al conjunto de entidades del sistema regulado por dicha institución y establece los "Requisitos mínimos de gestión implementación control de los riesgos relacionados con la tecnología informática, sistemas de información y recursos asociados para entidades financieras".

En el apartado referido a la Gestión de seguridad la comunicación establece que las entidades financieras deben considerar en su estructura organizacional un área para la protección de los activos de información, con el fin de establecer los mecanismos para la administración y el control sobre el acceso lógico y físico a sus distintos ambientes tecnológicos y recursos de información: equipamiento principal, plataforma de sucursales, equipos de departamentales, subsistemas o módulos administradores de seguridad de los sistemas de aplicación, sistemas de transferencias electrónica de fondos, base de datos, canales de servicios electrónicos, banca por Internet y otros.

Por la complejidad de la implementación de la Norma, sería conveniente que la Administración Pública definiera de manera sistemática el alcance el proyecto, principalmente en las áreas de CONTROL DE ACTIVOS Y SEGURIDAD DE LOS RECURSOS HUMANOS.

Adoptar una norma de gestión de la seguridad de la información garantiza la correcta consideración de los activos de información de una organización junto con el análisis de su vulnerabilidad y amenazas, también asegura que se establezcan controles de procedimiento y tecnológicos (gestión de accesos mediante registro de huellas, control de información por Internet y mails, controles sobre el software y las bases de datos). Mediante la norma estos procedimientos se desarrollan de manera ordenada, pautada y controlada; no como simples impulsos aislados, sino como un verdadero sistema que garantice la confidencialidad, integridad y acceso a la información de gestión, de dirección, y pública de la empresa.

A partir de la generalización de la tecnología de la información y la comunicación proliferaron las situaciones de riesgo para los sistemas de información, las bases de datos y los recursos de hardware, a los cuales denominamos activos de la información.

[6]B.O. 31.156,16/05/2007, corregida por la Comunicación C-48.583/2007 BCRA y actualizada por la Comunicación B-9042/2007 BCRA

Se tomó en cuenta la decisión administrativa 669/2004 que estableció que los organismos del Sector Público Nacional comprendidos en los incisos a) y c) del artículo 8º de la Ley Nº 24.156 y sus modificatorias deberán dictar o adecuar sus políticas de seguridad. Conformación de Comités de Seguridad en la Información. Funciones de los mismos y responsabilidades en relación con la seguridad.

Los comités deberán integrado al menos por un miembro del Directorio o autoridad equivalente, y el responsable máximo del área de Tecnologías Informática y sistemas.

III.- Infraestructuras críticas.

Infraestructuras Criticas son las infraestructuras estratégicas (es decir, aquellas que proporcionan servicios esenciales) cuyo funcionamiento es indispensable y no permite soluciones alternativas, por lo que su perturbación o destrucción tendría un grave impacto sobre los servicios esenciales". Se refiere tanto a empresas del sector TIC, agua, energía, industria nuclear, sistema financiero o transporte, ente otras.

Las Infraestructuras Criticas, son consideradas un conjunto de recursos, servicios, tecnologías de la información y redes, que en el caso de sufrir un ataque, causarían gran impacto en la seguridad, tanto física como económica, de los ciudadanos o en el buen funcionamiento del Gobierno de la Nación, siendo menester dictar medidas para la protección de tales infraestructuras.

Existen gran variedad de formas de ataque a los sistemas, con el fin de obtener esta información. Pero también existen políticas y paradigmas de seguridad actuales que nos permiten poner una brecha a estos ataques y proteger este recurso tan valioso.

Tomando en cuenta que el mundo contemporáneo se caracteriza por los profundos cambios originados en el desarrollo y difusión de las tecnologías de la información y la comunicación en la sociedad, las cuales se encuentran sustentadas en gran medida en el ciberespacio. Y que la utilización de las comunicaciones virtuales es un recurso que depende de la infraestructura digital, la cual es considerada como infraestructura crítica, entendiéndose ésta como imprescindible para el funcionamiento de los sistemas de información y comunicaciones, de los que a su vez dependen de modo inexorable, tanto el Sector Público Nacional como el sector privado, para cumplir sus funciones y alcanzar sus objetivos.

En el planteamiento de los instrumentos de planificación, se detectaron cambios estratégicos de protección que hasta ahora estaban enfocados principalmente a la seguridad física y que, ahora, como no podía ser de otra manera tienen un enfoque integral y actual de protección de las Infraestructuras críticas como conjunto de actividades destinadas a asegurar la funcionalidad, continuidad e integridad de las infraestructuras críticas con el fin de prevenir, paliar y neutralizar el daño causado por un ataque deliberado contra infraestructuras y garantizar la integración de estas actuaciones.

Se requiere un planteamiento de seguridad holístico (física, lógica, personal y operativa) de verdadera Seguridad Integral = Protección + Prevención donde la Gestión de riesgos debe ser su protagonista más importante y las soluciones pasen por la Evaluación de Impactos, establecimiento de Planes de contingencia, Planes de continuidad del negocio y de las operaciones y la determinación de los Sistemas y aplicaciones de garantía de alta disponibilidad.

En la Protección de Infraestructuras Críticas, es preciso estudiar los criterios que permitan determinar qué factores confieren carácter crítico a una infraestructura o elemento de infraestructura particular. Los criterios de selección deberían basarse en conocimientos sectoriales y generales. Pueden definirse tres factores de identificación de una infraestructura crítica potencial:

• Alcance - la pérdida de un elemento de infraestructura crítico se mide por el tamaño del área geográfica que pudiera verse afectada por su pérdida o indisponibilidad.

• Magnitud - el grado del impacto o de la pérdida puede evaluarse como nulo, mínimo, moderado o principal. Entre los criterios que podrían utilizarse para evaluar la magnitud potencial se encuentran los siguientes:

(a) impacto público (cantidad de población afectada, pérdidas de vidas, enfermedades, lesiones graves, evacuación);

(b) económico (efecto PIB, volumen de pérdida económica y/o degradación de productos o servicios);

(c) ambiental (impacto en el lugar y sus alrededores);

(d) interdependencia (con otros elementos de infraestructura críticos).

(e) político (confianza en la capacidad de las administraciones públicas);

Efectos en el tiempo - estos criterios determinan en qué plazo (tiempo) la pérdida de unelemento podría tener un impacto. En caso de sufrir un ataque, las estructuras críticas, causarían gran impacto en la seguridad, tanto física como económica del país. Este impacto se mide según unos criterios horizontales que determinan la

criticidad de una infraestructura. Se han establecido tres: el número potencial de víctimas, el impacto económico y el impacto público**.**

En la República Argentina se enuncia que El Programa Nacional de Infraestructuras Críticas de Información y Ciberseguridad (ICIC) tiene como finalidad impulsar la creación y adopción de un marco regulatorio específico que propicie la identificación y protección de las infraestructuras estratégicas y críticas del Sector Público Nacional, los organismos interjurisdiccionales y las organizaciones civiles y del sector privado que así lo requieran, y la colaboración de los mencionados sectores con miras al desarrollo de estrategias y estructuras adecuadas para un accionar coordinado hacia la implementación de las pertinentes tecnologías.

También impulsa una Encuesta Nacional sobre Acceso y Uso de Tecnologías de la Información y la Comunicación (ENTIC) en Hogares y Personas, que permite contar con información desde la perspectiva de los usos y accesos de los hogares y de las personas a dichas tecnologías en la Argentina. La ENTIC se administró a todos los hogares para la Encuesta Anual de Hogares Urbanos (EAHU), cuya estimación se extiende al total de la población residente en hogares particulares urbanos, en localidades de 2.000 o más habitantes.

Mediante el Programa Nacional de Infraestructura Crítica de Información y Ciberseguridad (ICIC) e Internet Sano

- Se promoverá la concientización de la protección de las infraestructuras críticas de información y la ciberseguridad dentro de las dependencias del Sector Público Nacional, brindando asistencia técnica a los organismos nacionales, provinciales y municipales que lo requieran.
- Se actualizará la Estrategia Nacional ICIC.
- Se dictarán talleres y charlas técnicas sobre Ciberseguridad.
- Se formularán ejercicios de respuesta a incidentes.
- Se creará la Política de Seguridad de la Información.
- Se generarán nuevos contenidos para concientización de la ciudadanía.
- Se desarrollarán exposiciones y conferencias de concientización.
- Se concertarán alianzas público-privadas para la creación y difusión de contenidos.

Organismo responsable: Subsecretaría de Tecnologías de Gestión. Jefatura de Gabinete de Ministros. Fecha tentativa: diciembre 2013

IV. Propuesta para implementar en la UNSJ.

La Universidad Nacional de San Juan (U.N.S.J.) ha desarrollado una plataforma tecnológica a través de la cual se registra, procesa, transmite y almacena información mediante diferentes activos de Información, que permiten interactuar con la comunidad académica, ciudadanía en general y el personal universitario de todo el país, y como además se reconoce que la información que posee es un bien estratégico para sus fines, por lo que se requiere que sea protegida también su obtención, procesamiento, transmisión y almacenamiento. Considerando que la Resolución 580/2011 crea, el "Programa Nacional De Infraestructuras Criticas De Información y Ciberseguridad" en el marco de lo establecido la Ley de Ministerios (t.o. Decreto Nº 438/92), a fin de impulsar la creación y adopción de un marco regulatorio específico que propicie la identificación y protección de las infraestructuras estratégicas y críticas del Sector Público Nacional, los organismos interjurisdiccionales y las organizaciones civiles y del sector privado que así lo requieran. Dado que las universidades nacionales como órganos académicos deberían adherirse a lo previsto por la resolución 580/2011 que se creó, en el ámbito de la Oficina Nacional de Tecnologías de Información de la subsecretaria de Tecnologías de Gestión de la Secretaria de Gabinete de la Jefatura de Gabinete de Ministros, el "Programa Nacional De Infraestructuras Criticas De Información y Ciberseguridad" en el marco de lo establecido la Ley de Ministerios (t.o. Decreto Nº 438/92), se impone para la U.N.S.J. la obligación de establecer una Política de Seguridad que fije las directrices generales que oriente la materia de seguridad dentro de cada Unidad.

Para ello se crearía un comité que será el responsable máximo del área de Tecnologías Informática y sistemas. Y que deberá dictar o adecuar sus políticas de seguridad de la Universidad Conformación de Comités de Seguridad en la Información. Funciones de los mismos y responsabilidades en relación con la seguridad.

El encargado de seguridad de los activos de información debería establecer un organigrama de funciones, determinando la cantidad de miembros de la Unidad, cargos y funciones (discriminando quienes son los administradores de seguridad y/o miembros del comité). Tendría además a su cargo el desarrollo y actualización de las políticas de seguridad y controlar su implementación, utilizando como referencia las

Normas IRAM ISO 27.001[7] y 27002[8] y su antecedente 17799) debido a que en base a ellas se definieron los requisitos para el sistema de gestión de seguridad (SGSI) propuesto.

Para ello sería necesario para ello adoptar las medidas para la protección de la Infraestructuras críticas donde se fije entre otras cosas, la necesidad de que para esta gestión se fije:

- Un Plan de Seguridad del Operador (PSO)
- Un Plan de Protección Especifico (PPE) para cada una de las infraestructuras que haya sido identificada como critica por la Secretaria de para la Protección de Infraestructuras Criticas.

El encargado de seguridad debería recoger los contenidos mínimos que deben articular estos planes, y desarrollar un nuevo documento en el que se describe el modo de abordar la implantación de las medidas y más tarde reflejarlo en dichos planes.

Se entiende por incidente de seguridad todo incidente que impida el normal funcionam8ento de los activos de información y que afecte la Seguridad Informática.

La gestión de incidentes de Seguridad tiene por objeto restaurar la operación normal de los sistemas con tanta rapidez como sea posible y mitigar el impacto adverso a sus procesos, asegurando así que se mantenga debidamente la confidencialidad, integridad y disponibilidad de la información de la U.N.S.J.

El comité de Seguridad Informática definirá las pautas a seguir en la gestión de incidentes de seguridad, lo que deberán ser implementados por el Departamento de Sistemas de la U.N.S.J. bajo la coordinación del encargado de Seguridad de los activos de información.

Además deberá elaborar una Guía aplicativa del sistema de seguridad utilizando como apoyo la Norma IRAM ISO IEC 27001.

Se propone un sistema de gestión de Incidentes de seguridad porque es la forma en que la Organización dirige y controla aquellas actividades asociadas a la seguridad. De una manera más amplia debería contar con dos grandes apartados alineados con la estructura del Plan de Seguridad del Operador (PSO) y del Plan de protección Específico (PPE):

- Un capítulo dedicado al análisis de riesgos, uno de los aspectos principales de los mencionados planes.

Dos capítulos dedicados a recoger las medidas de seguridad lógicas y físicas que se deberán:

- Construir y proponer una normativa de seguridad aplicable al entorno local mediante la
- Implementación de un Equipo de Respuesta a Incidentes para la Universidad Nacional de San Juan.

La Metodología propuesta para el Manejo de Incidentes CSIRT[9] – UNSJ, básicamente se ha estructurado de acuerdo al siguiente esquema:

Cada uno de los pasos propuestos se describe a continuación:

1. Preparación y Protección

La fase de preparación consiste, principalmente, en la implementación de un equipo de Respuesta a Incidentes de Seguridad Informática (CSIRT), las actividades propuestas, que se deben realizar en esta fase son:

- Planificación del CSIRT – UNSJ
- Implementación del CSIRT – UNSJ
- Evaluación y funcionamiento del CSIRT
- Lecciones Aprendidas.

A más de definir un proceso de implementación de un equipo de CSIRT, es importante tomar en cuenta la Protección de la infraestructura de la Universidad para de esta manera asegurar que los sistemas, redes y aplicaciones tengan un nivel de seguridad adecuado. Las actividades de esta fase se realizan en conjunto con

[7] En Argentina es IRAM, como organismo nacional de normalización, quien la estudia a través del Subcomité de Seguridad de la Información y la adopta como IRAM-ISO/IEC 27001. Se publica bajo el nombre Tecnología de la información. Sistemas de gestión de la seguridad de la información (SGSI). Requisitos, difundiéndola en la región a través de cursos y seminarios.

[8] Aprobada y consensuada por el IRAM (Instituto de Normalización Argentino) en el año 2002

[9] Un **Equipo de Respuesta ante Emergencias Informáticas** (**CERT**, del inglés **Computer Emergency Response Team**) es un centro de respuesta a incidentes de seguridad en tecnologías de la información. Se trata de un grupo de expertos responsable del desarrollo de medidas preventivas y reactivas ante incidencias de seguridad en los sistemas de información. Un CERT estudia el estado de seguridad global de redes y ordenadores y proporciona servicios de respuesta ante incidentes a víctimas de ataques en la red, publica alertas relativas a amenazas y vulnerabilidades y ofrece información que ayude a mejorar la seguridad de estos sistemas.

el área de Seguridad, pero básicamente el área de Manejo de incidentes tiene a su cargo la prevención de ataques, y si estos suceden mitigar el impacto. El área de seguridad realiza actividades de protección, en cuanto a configuraciones y garantiza la infraestructura informática de la Universidad.

2. Detección de Incidentes de Seguridad

Esta fase está compuesta de varias actividades, tales como: detección de incidentes, análisis inicial y documentación del incidente y tiene como objetivo la búsqueda de toda posible señal de ocurrencia de un incidente. Todas las actividades e información generada en esta fase en enviada al proceso de Triage, haciendo uso de los reportes establecidos.

La Detección de incidentes es un proceso que permite identificar las actividades inusuales que pueden comprometer la misión del CSIRT, consiste en la detección y evaluación de posibles incidentes, determinar si un incidente ha ocurrido, y de ser así, el tipo, extensión y magnitud del problema.

Estas actividades se pueden identificar de manera reactiva y proactiva.

Los incidentes se pueden detectar a través de muchos medios tales como: IDS basados en red

(NIDS) y en host (HIDS), software antivirus, software de control de integridad de archivos, sistemas de monitoreo de red, analizadores de logs, etc.

Los incidentes también pueden ser detectados por medios manuales, tales como reportes de incidentes de usuarios.

En el proceso de detección están involucrados: Encargado de Seguridad, CSIRT-UNSJ, Gestión de

Servicios TI, Infraestructura de TI, usuarios que han sido víctimas de algún ataque y otras áreas, incluye los siguientes aspectos:

- Señales de un incidente
- Detección de incidentes mediante la utilización de herramientas
- Detección de incidentes mediante el reporte de terceros.

Señales de un Incidente:

En el proceso de detección, la información sobre potenciales incidentes, vulnerabilidades, información de seguridad informática o de manejo de incidentes, puede ser obtenida de dos maneras:

- Detección Reactiva

Un incidente puede haber ocurrido o estar ocurriendo en este momento, puede ser detectado de varias maneras:

- El antivirus detecta que un equipo está contaminado con algún tipo de virus.
- Incidentes en el servidor web
- Envío de alertas y notificaciones por parte de otras organizaciones.
- Detección Proactiva
- Monitoreo de Red
- Escaneo de vulnerabilidades
- Investigación
- Análisis de Riesgos

La detección de incidentes es un proceso que permite saber si el sistema está en peligro o si los servidores corren el riesgo de detener sus servicios.

Esta actividad va de la mano con la detección proactiva, se debe tomar en cuenta el personal que se encargue del monitoreo y detección de actividad sospechosa, análisis de logs, uso de software de detección de intrusos, para cada una de estas actividades se deben tomar en cuenta los procesos establecidos en el Área de Seguridad para estas actividades.

Todos los datos analizados y los considerados sospechosos se envían al proceso de Triage.

Detección de incidentes mediante el reporte de terceros va de la mano con la detección reactiva, el usuario notifica del incidente al área de Gestión de Servicios, si el incidente se encuentra en la base de conocimiento de esta área, es atendido por ellos, caso contrario se envía el reporte del incidente al equipo CSIRT – UNSJ, en donde, primeramente se verifica que sea un incidente de seguridad.

Los incidentes que se envían al CSIRT-UNSJ y los que se atienden, son los que constan en la Categorización de incidentes del CSIRT-UNSJ.

- Análisis de Incidentes de Seguridad

En este proceso se busca analizar cada reporte de incidentes presentado, tanto por los usuarios y por los reportes obtenidos de las herramientas utilizadas, con la finalidad de verificar si realmente se trata de un incidente de seguridad, o son falsos positivos.

Se debe recalcar que el equipo CSIRT debe trabajar rápidamente en el análisis y validación de los incidentes, todas las acciones realizadas deben ser documentadas.

Uno de los mecanismos que sirve de soporte para Detección es la documentación del incidente, se han definido varios formatos para el reporte y respuesta de incidentes y vulnerabilidades, el uso de reportes ayuda a:

- Proveer información completa de un incidente al equipo
- Organizar la información recibida
- Priorizar reportes

La información que se solicita en el reporte de incidente es:

- Información de contacto
- Fecha de reporte
- Sistemas afectados
- Descripción del incidente
- Observaciones

A más de los reportes de incidentes, parte de la documentación incluye un documento en el que se detalla el cómo los usuarios deben realizar el reporte de los incidentes al equipo, etc.

Adicional al reporte de incidente enviado por el usuario, por parte del Equipo CSIRT-UNSJ se debe enviar un documento de respuesta a incidentes, en el que se detalle la información relativa a la atención y respuesta del incidente reportado, dependiendo del tipo de incidente, esta información será enviada a autoridades, y personal que requiera de esta información. (Esta sección, se detalla en la fase de la respuesta a incidentes).

En cuanto a los recursos humanos, es necesario que el plantel de empleados sea idóneo para la realización del trabajo de la Organización, pero además debe definir y comunicar sus funciones y responsabilidades.La misma organización debe establecer las necesidad información y facilitar y evaluar la eficacia de la formación y de esto debe haber evidencia (Es decir, se exige un registro de capacitación del personal y en lo posible la formación de una tablero de comando al efecto).

También se debe sensibilizar a toda la organización sobre la importancia de la capacidad humana de la Organización descansa sobre la formación que da a todos sus empleados. La organización dispone un potencial que debe ser aprovechado para poder subsistir y este es el potencial humano para ello se debe implantar los siguientes aspectos motivación, adiestramiento y Comunicación.Implantar en las infraestructuras críticas para mejorar los niveles de protección integrales.

Se propone crear un CSIRT dentro de la UNSJ, ello es un proceso que involucra un cambio estructural, organizacional y desde luego requiere de mucho esfuerzo y compromiso a todos los niveles. Sin duda un CSIRT viene a darle un gran valor a la organización ya que provee un punto de contacto único para afrentar, resolver y proponer en el campo de las nuevas tecnologías.

La protección del ciberespacio requiere de una organización que sirva de centro nacional decoordinación para asegurar y proteger el ciberespacio, cuya misión incluye vigilancia, alerta,respuesta y recuperación, con la colaboración de las entidades gubernamentales en los ámbitosnacional, estatal y local, el sector privado, el sector académico y la comunidad internacional.

Los principales objetivos que debe cubrir este centro nacional son:

Desarrollar un sistema nacional de seguridad y respuesta ante incidentes cibernéticos paradetectar, prevenir, responder y recuperarse de incidentes en el ciberespacio.

- Establecer un centro de coordinación para la gestión de incidentes cibernéticos que reúnen los elementos críticos del gobierno y los elementos esenciales de las infraestructuras de los operadores y los proveedores para reducir el riesgo y la gravedad de los incidentes.
- Participar en la vigilancia, alerta y mecanismos de intercambio de información.
- Elaborar y probar los planes de respuesta de emergencia, procedimientos y protocolospara asegurar que los colaboradores gubernamentales y no gubernamentales puedanfomentar la confianza ypuedan coordinarse de manera eficaz en caso de crisis.

La implementación del CSIRT-UNSJ permitirá:

- Contar con un equipo capacitado para la atención de incidentes brindando servicios proactivos, reactivos y de aseguramiento de la calidad en temas de seguridad de la información.
- Concientizar a la comunidad universitaria y usuarios finales sobre los riesgos y beneficios del uso de internet, pero, sobre todo de la importancia de tomar en cuenta las medidas de seguridad adoptadas en la Universidad.

Por otra parte como objetivo fundamental el construir y proponer una normativa de seguridad aplicable al entorno local, la implementación del equipo permitirá compartir la experiencia y resultados obtenidos con otras universidades.

Los beneficiarios de la implementación de este proyecto serán principalmente la UNSJ y con la implementación del proyecto:

- Otras Universidades, con la finalidad de profundizar y proponer temas de investigación.
- Organizaciones públicas y privadas que mantienen Equipos de Seguridad.
- Empresas y profesionales que brindan servicios de atención de incidentes.
- Participación efectiva de todos los intervinientes en el proyecto.

Hasta la fecha la participación de los integrantes del equipo se ve reflejada en el cumplimiento de las actividades, estas se han realizado acorde a lo planificado.

Se ha involucrado a personal de Marketing para la publicidad del CSIRT.

Se han realizado reuniones de Coordinación

Se mantuvieron reuniones con:

- Grupo de seguridad para la presentación de la metodología para el manejo de incidentes.
- Grupo de Administradores para comunicarles la existencia del CSIRT-UNSJ, uso de reportes de incidentes y políticas.

Se realizaran actividades de Difusión mediante emisión de boletines con tips de seguridad para usuarios, los mismos que se emitirán mensualmente. Como estrategia de marketing, durante el primer mes se emitirán boletines semanalmente, con el objetivo de posicionar el CSIRT-UNSJ en la comunidad universitaria.

- En conjunto con el área de marketing de la UNSJ se está preparando unaestrategia de comunicación para realizar la difusión del CSIRT-UNSJ, lo que incluye boletines informativos, notas periodísticas, etc.
- Emisión de un boletín sobre el CSIRT-UNSJ al área de marketing para que sea difundido a nivel interno en la UNSJ

Finalmente la guía deberá incluir un Anexo dedicado a la protección de los sistemas de monitorización y control de procesos e infraestructuras, dada su relevancia en este tipo de infraestructuras, así como un apartado en el que se recogen, de manera exhaustiva, todas las referencias utilizadas.

V. Conclusiones

La seguridad de la infraestructura digital se encuentra expuesta a constantes amenazas, que en caso de materializarse pueden ocasionar graves incidentes en los sistemas de información y comunicaciones, por lo que resulta imprescindible adoptar las medidas necesarias para garantizar el adecuado funcionamiento de las infraestructuras críticas.Los riesgos se han incrementado y sofisticado y hay una demanda de mayor eficacia que exige nuevas respuestas que requieren tecnología, eficacia y calidad. Eficacia y calidad que deben ser percibidas por el usuario.

La propuesta para gestión integral de la seguridad de la información diseñada para la protección de la información de las organizaciones locales, basada en las tendencias, normas de seguridad y estándares actuales y la creación y adopción de un marco regulatorio que favorecerá la identificación y protección de las infraestructuras estratégicas y críticas de la U.N.S.J. , promoverá la colaboración entre los distintos sectores y propiciara el desarrollo de estrategias y estructuras adecuadas para la protección de los activos de información de las organizaciones locales.

El proyecto de Creación e Implementación de un CSIRT Académico para la Universidad Nacional de San Juan, tiene como objetivo fundamental, el construir y proponer una normativa de seguridad aplicable al entorno local, la implementación del equipo permitirá compartir la experiencia y resultados obtenidos con otras universidades Nacionales con el objetivo de proponer la creación de una red nacional de CSIRTs académicos y contribuir así a la investigación y desarrollo de metodologías y buenas prácticas que permitan mejorar la seguridad de las redes.

Así hemos de concluir en que, sin duda hoy la responsabilidad y respuesta de una única Seguridad con mayúscula, integral e integrada, pública y privada, es estrictamente necesaria e irreversible.Por todo ello, para un adecuado presente y futuro hay que integrar el sistema de gestión de la Seguridad Pública y la Seguridad Privada hacia una nueva visión común y especial cultura de seguridad sobre la base de las amenazas complejas y la interdependencia e incrementar los recursos de análisis y liberarlos de viejas patologías y rigideces para desarrollar el esquema de Gestión Integral e Integrada de la Seguridad.

VI.-Referencias

31"Evaluación y Seguridad de un Sistema de Información", por José Alfredo Jiménez
http://www.monografias.com/trabajos/seguinfo/seguinfo.shtml

32. "Administración segura de la información: Una experiencia de vinculación entre un ente del estado provincial y la U.N.P.A". Javier Díaz, L.I.N.T.I. – Universidad Nacional de La Plata
http://www.ing.unp.edu.ar/wicc2007/trabajos/ISBD/066.pdf

33. "Propuesta para un modelo de gestión de documentos electrónicos de archivo en la administración pública"- documento de trabajo elaborado por Carlos Alberto Zapata y Nelson Javier Pulido para el comité de gestión de documentos del sistema nacional de archivos

34.http://www.slideshare.net/scarchivistas/propuesta-para-un-modelo-de-gestin-de-documentos-electrnicos-de-archivo-en-la-administracin-pblica

Gómez Vieites, Álvaro (2007). Enciclopedia de la Seguridad Informática.

3.5Altmark, Daniel Ricardo y Molina Quiroga Eduardo. Tratado de Derecho Informático. Tomo III. Publicado por la Ley año 2012

Model Design for a Reduced Variant of a Trivium Type Stream Cipher

Antonio Castro Lechtaler[1], Marcelo Cipriano[1], Edith García[2], Julio Liporace[2], Ariel Maiorano[2], Eduardo Malvacio[2],
Escuela Superior Técnica "Gral. Div. Manuel N. Savio", Facultad de Ingeniería
Instituto de Educación Superior del Ejército Argentino

[1]{acastro, marcelocipriano}@iese.edu.ar;
[2]{edithxgarcia, jcliporace, maiorano, edumalvacio}@gmail.com

Abstract. We analyze the family of stream ciphers N-viums: Trivium and Bivium. We present the Trivium algorithm and its variants. In particular, we study the NLFSRs used in these generators, their feedback functions and their combination. Two reduced variants of these models are presented, labeled Toys. Finally, we delve into the open problems ingrained in these cryptosystems.

Keywords: LFSR, NLSFR, Trivium, Bivium, Trivium-Toy, Bivium-Toy.

1 Introduction

The revolution of communications and technology has taken cryptology from the military and diplomatic realm into everyday life.
E-mailing, home banking, user authentication in social networks, mobile communications, and wireless technology have increased the requirements for confidentiality while data is transferred via insecure channels.

Some ciphering systems meet the requirements to protect data satisfactorily. However, they do not meet the increasing demand for higher transfer rates.

Because of the resources used and the processing power required, the existing algorithms lag behind the increasing needs for data transfer security.

Stream ciphers may prove suitable to use in portable devices. Their hardware adaptability turns them into feasible solutions, responding to the increasing demand and high transfer rate standards.

1.1 Stream Ciphers.

A perfect cryptosystem entails the capacity for an algorithm to cipher a message which can be deciphered only by the intended receiver.

Vernan and Mauborge created such a system in 1917 at the AT&T labs. In their design, the required key is as long as the length of the message. Both, transmitter and receiver must have the key which must be destroyed after use. Otherwise, security is jeopardized.

Because of this feature, the system is known as One-Time-Pad. The key must be random and is used for both processes: ciphering and deciphering. Hence, users need to share it at both ends. Cryptosystems under this particular secret key configuration belong to a class known as symmetric-key algorithms.

In 1949, Shannon demonstrated the invulnerability of this system by satisfying the requirements for perfect secrecy established by the rising field of Information Theory.

Nonetheless, two weaknesses become apparent, not in the algorithm itself, but in its application. On one hand, a problem arises in the generation of the secret key; and, on the other, in the security of key distribution.

A possible solution is to find a deterministic procedure to generate the key. Such a key would not be random, but pseudorandom, and shall meet additional requirements to be considered secure.

1.2 LFSRs and Non-LFSRs

Currently, *Linear Feedback Shift Registers* (LFSRs) are used extensively to generate pseudorandom sequences with controlled period and linear complexity.

Research on LFSRs began in the 60s [6] and continued through several years. A significant number of results and applications have been produced: algorithm design, error control codes, and linear complexity analysis of binary sequences with the Berlekamp-Massey algorithm [7].

Because of their linearity, LFSRs alone are insecure. It is widely known that, when *2n-1* consecutive bits of an outbound sequence are known, it becomes predictable. Attempts to add linear complexity by combining LFSRs **w**ith, among other things, nonlinear functions have not met the desired standards yet.

Nonlinear Feedback Shift Registers (NLFSRs), a generalization of their linear counterparts, have been relegated for a long time. While LFSR theory is robust and well understood, many fundamental problems with NLFSRs remain unanswered.

One such problem is the determination of the period of outbound sequences in NLFSRs. In recent years, research has focused on nonlinear registers and stream ciphers using NLFSRs in some form. This is the case for the class TRIVIUM [1][2], BIVIUM [10].

Our research focuses in the development of a new family of the TRIVIUM-BIVIUM stream cipher class, designated as *Toys*.

In our Toys, in which the sizes of the NLFSRs are reduced significantly, we have modified their taps while maintaining the original design principles.

With these models, observation in a constrained environment may foster more realistic research projects, as well as allow researchers to compare results within smaller samples and to conduct tests in a reduced space.

In the future, the Toy family may help contribute in the development of a solid algebra involving NLFSRs, in particular for generators of the TRIVIUM-BIVIUM class.

2. FSR Overview

An n-bit *feedback shift register* (FSR) is an n-bit length register with a feedback function:

$$f:\{0,1\}^n \rightarrow \{0,1\} \quad \textbf{(1)}$$

where the feedback bit (at the tap positions of the register) or the output bit is of the form:

$$x_{n+t} = f(x_{n-1+t}, x_{n-2+t}, \ldots\ldots x_t) \ (t \geq 0) \quad \textbf{(2)}$$

For each step *t*, the register bits shift one position to the right and the taps are fed into the function and become the bit input for the following step. The *n* bits of the register constitute the state of the register at step *t*. The initial state is defined when *t=0*. The period of a FSR is the length of the largest cycle generated by the output sequence of the register.

If the feedback function is linear, i.e.:

$$f(x_{n-1}, x_{n-2}, \ldots\ldots x_0) = c_0x_0 + c_1x_1 + c_2x_2 + \ldots\ldots\ldots + c_{n-1}x_{n-1} \ (c_i \in \{0,1\}) \quad \textbf{(3)}$$

we say that the registry is an **LFSR (Linear Feedback Shift Register)**. Otherwise, with a nonlinear feedback function, we have a **NLFSR (Nonlinear Feedback Shift Register)**.

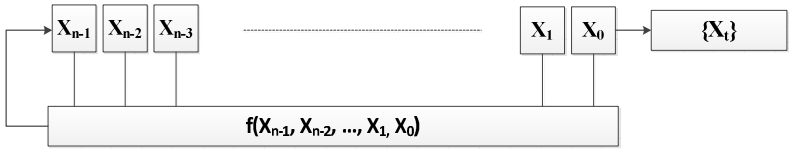


Fig.1: n-bit FSR Structure.

In the LFSR case, when the coefficients c_i belong to a primitive polynomial, the LFSR output sequence has a maximum length of 2^n-1, regardless of the chosen initial (non-trivial) state. The LFSR output sequences of maximum length are called *maximal sequences or m-sequences* [6]. If *2n-1* output bits of an n-length LFSR are known, then the sequence becomes predictable using the Berlekamp-Massey algorithm [10].

NLFSRs are more robust to algebraic attacks. However, no systematic and efficient method is known to construct secure NLFSRs [3][4]. Furthermore, for a given nonlinear feedback function, it is difficult to predict the period of the output sequence.

A **stream cipher** is a symmetric ciphering system which takes a sequence of plaintext and a secret key, and operates on the plaintext, generally bit by bit with the **key bit stream,** generated by the secret key and the algorithm.

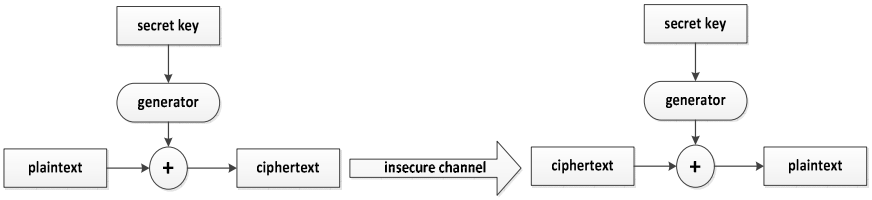


Fig.2 Stream Cipher Example

The key bit stream must meet certain cryptologic security conditions, i.e.: the length of the sequence and the linear complexity must be sufficiently large, and the binary sequence must satisfy a series of pseudo-random tests [6].

3. Trivium and Bivium

The stream algorithm TRIVIUM was designed by Christophe De Cannière and Bart Preneel. It was selected as a finalist algorithm in the e-STREAM Project [5]. It was designed to generate at least 2^{64} bits with the use of an 80-bit secret key and an initialization vector (IV) of also 80 bits.

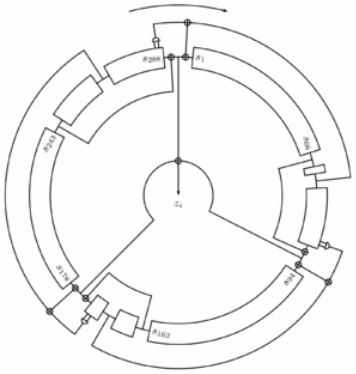


Fig.3: Trivium algorithm

It consists of three combined NLFSRs. The first register controls the second, the second controls the third, and this last register controls the first.

Fig.4: Trivium-Like Structure

The core idea behind the design focuses on using the principles of block cipher construction in order to create equivalent components in stream ciphers.

The output consists of three combined non-linear shift registers of length *93*, *84*, and *111*, and where specific positions are selected to obtain a key bit stream. Whereas no efficient attack has been encountered to break this generator so far [8][9], its period remains undetermined and an open research problem.

A complete description is given by the following simple pseudo-code:

INPUT: $s_0, s_1, \ldots, s_{287}$ initial state, integer n, $s_i \in \{0,1\}$.

OUTPUT: binary sequence $\{k_t\}$

1. Initialization

$t_1 \leftarrow s_{65} \oplus s_{92}$
$t_2 \leftarrow s_{161} \oplus s_{176}$
$t_3 \leftarrow s_{242} \oplus s_{287}$

2. While (t<n) do the following:

2.1 $k_t \leftarrow t_1 \oplus t_2 \oplus t_3$
2.2 $t_1 \leftarrow t_1 \oplus s_{90} \otimes s_{91} \oplus s_{170}$
$t_2 \leftarrow t_2 \oplus s_{174} \otimes s_{175} \oplus s_{263}$
$t_3 \leftarrow t_3 \oplus s_{285} \otimes s_{286} \oplus s_{68}$
2.3 $(s_0, s_1, \ldots, s_{92}) \leftarrow (t_3, s_0, \ldots, s_{91})$
$(s_{93}, s_{94}, \ldots, s_{176}) \leftarrow (t_1, s_{93}, \ldots, s_{175})$
$(s_{177}, s_{178}, \ldots, s_{287}) \leftarrow (t_2, s_{177}, \ldots, s_{285})$

3. Return $\{k_t\}$

Note that $\oplus$ is the XOR operation and $\otimes$ the AND operation.

BIVIUM was designed by Hårvard Raddum to obtain a reduced sized version of TRIVIUM. It consists of two combined NLFSRs (while TRIVIUM has three) of lengths *93* and *84*.

Despite the improved security under specific attacks granted by this model, the results are not entirely satisfactory.

4. The Toy Model

We present reduced variants of TRIVIUM and BIVIUM algorithms as a strategy to tackle the open problems discussed and the mathematical theory behind the behavior of NLFSRs. The reduced models (decimated by 3) are based on previous work by Yun Tian et al, who developed an extended model of the TRIVIUM structure [11]. We have named these models Toys, considering they are *miniatures* of the originals.

It is noted that every reduction of a model focuses on a quest for simplicity in its mathematical study and it is not meant to be used in operative information security environments.

We assume the following:

A1) *Property invariance after size reduction*: the reduced size structure of the models maintains the mathematical properties of the original model.

A2) *Computational complexity reduction:* The reduction in size contributes to a reduction of the problem, making the model more manageable under computational as well as algebraic considerations.

A3) *Property invariance after size increase:* In the case of identified patterns in the behavior and mathematical properties in the reduced model, they may be extrapolated to the original model.

These assumptions need to hold throughout the entire research. In case one of them does not hold or inconsistences among them are encountered, the procedure presented here ought to be revised.

4.1. Trivium-Toy

The model consists of three NLFSRs *X, Y,* and *Z* of lengths ***31, 28*** and ***37*** with the following states:

$$\begin{aligned} &X(31): && X_0, X_1, \ldots\ldots\ldots\ldots, X_{30} \\ &Y(28): && Y_0, Y_1, \ldots\ldots\ldots\ldots, Y_{27} \\ &Z(37): && Z_0, Z_1, \ldots\ldots\ldots\ldots, Z_{36} \end{aligned} \quad \textbf{(4)}$$

Being the feedback of each register, i.e. the bit input in each:

$$\begin{aligned} &X_0: Z_{21} \oplus Z_{36} \oplus Z_{35} \otimes Z_{34} \oplus X_{22} \\ &Y_0: X_{21} \oplus X_{30} \oplus X_{29} \otimes X_{28} \oplus Y_{25} \\ &Z_0: Y_{22} \oplus Y_{27} \oplus Y_{26} \otimes Y_{25} \oplus Z_{28} \end{aligned} \quad \textbf{(5)}$$

and the key bit stream:

$$K_t: X_{21} \oplus X_{30} \oplus Y_{22} \oplus Y_{27} \oplus Z_{21} \oplus Z_{36} \quad \textbf{(6)}$$

Also, the cipher of the plaintext with the key bit stream is:

$$C_t = P_t \oplus K_t \quad \textbf{(7)}$$

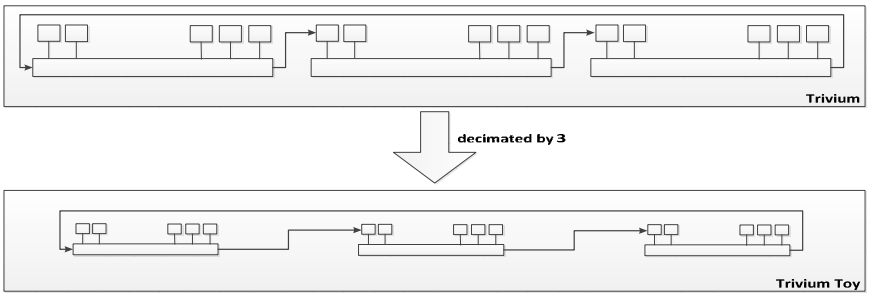


Fig.5 Trivium vs Trivium Toy

Pseudo-code of the Trivium is changed to a reduced form as follows:

INPUT: $s_0, s_1, \ldots, s_{95}$ initial state, integer n, $s_i \in \{0,1\}$.

OUTPUT: binary sequence $\{k_t\}$

1. Initialization.

   $t_1 \leftarrow s_{21} \oplus s_{30}$
   $t_2 \leftarrow s_{53} \oplus s_{58}$
   $t_3 \leftarrow s_{80} \oplus s_{95}$

2. While (t<n) do the following:

   2.1 $k_t \leftarrow t_1 \oplus t_2 \oplus t_3$

   2.2 $t_1 \leftarrow t_1 \oplus s_{28} \otimes s_{29} \oplus s_{55}$
   $t_2 \leftarrow t_2 \oplus s_{56} \otimes s_{57} \oplus s_{87}$
   $t_3 \leftarrow t_3 \oplus s_{93} \otimes s_{94} \oplus s_{22}$

   2.3 $(s_0, s_1, \ldots, s_{30}) \leftarrow (t_3, s_0, \ldots, s_{29})$
   $(s_{31}, s_{32}, \ldots, s_{58}) \leftarrow (t_1, s_{31}, \ldots, s_{57})$
   $(s_{59}, s_{60}, \ldots, s_{95}) \leftarrow (t_2, s_{59}, \ldots, s_{94})$

3. Return $\{k_t\}$

4.2. Bivium-Toy

The model consists of two **NLFSRs** *X, and Y* of lengths ***31*** and ***28*** respectively with the following states:

$$\begin{array}{ll} X(31): & X_0, X_1, \ldots\ldots\ldots\ldots, X_{30} \\ Y(28): & Y_0, Y_1, \ldots\ldots\ldots\ldots, Y_{27} \end{array} \quad \textbf{(8)}$$

Being the feedback of each register:

$$\begin{array}{l} X_0: Y_{22} \oplus Y_{27} \oplus Y_{26} \otimes Y_{25} \oplus X_{22} \\ Y_0: X_{21} \oplus X_{30} \oplus X_{29} \otimes X_{28} \oplus Y_{25} \end{array} \quad \textbf{(9)}$$

and the key bit stream:

$$K_t: \qquad X_{21} \oplus X_{30} \oplus Y_{22} \oplus Y_{27} \quad \textbf{(10)}$$

The cipher process is the same as detailed in formula **(7).**
Pseudo-code of this reduced cipher is given below:

INPUT: $s_0, s_1, \ldots, s_{58}$ initial state, integer n., $s_i \in \{0,1\}$.

OUTPUT: binary sequence $\{k_t\}$

1. Initialization.

$t_1 \leftarrow s_{21} \oplus s_{30}$
$t_2 \leftarrow s_{53} \oplus s_{58}$

2. While (t<n) do the following:

2.1 $k_t \leftarrow t_1 \oplus t_2$

2.2 $t_1 \leftarrow t_1 \oplus s_{28} \otimes s_{29} \oplus s_{55}$
$t_2 \leftarrow t_2 \oplus s_{56} \otimes s_{57} \oplus s_{22}$

2.3 $(s_0, s_1, \ldots, s_{30}) \leftarrow (t_2, s_0, \ldots, s_{29})$
$(s_{31}, s_{32}, \ldots, s_{58}) \leftarrow (t_1, s_{31}, \ldots, s_{57})$

3. Return $\{k_t\}$

5. Conclusions

In this article we present the class of Trivium-Bivium random sequence generators using non-linear shift registers (NLFSR).

Because of their size, several research problems remain unanswered: patterns of behavior, algebraic properties, period lengths, and weak keys among others.

Under this framework, we present reduced sized variants of these generators for research and applications in cryptology, laying out the formulae of the feedback functions as well as the key bit streams. We assume that the properties identified in the reduced sized models would remain invariant in the original ones.

6. Further research.

The Toy family may foster additional research in the following areas:
- Search for length of the period or cycles.
- Distribution of taps and their changes to determine algebraic properties and personalization of N-viums.
- Algebraic analysis of the non-linear functions used in the models.
- Search for possible weak keys.

7. Acknowledgements

The financial support provided by Agencia Nacional para la Promoción Científica y Tecnológica (Project PICTO 11- PICTO 11-18621) is gratefully acknowledged.

8. References

1. De Canniére, C. and Preneel, B. "*TRIVIUM A Stream Cipher Construction Inspired by Block Cipher Design Principles*". In Workshop on Stream Ciphers Revisited, (2006).
2. De Canniére, C. and Preneel, B. "*TRIVIUM Specifications*". eSTREAM, ECRYPT Stream Cipher Project, Report. (2008).
3. Dubrova, E. "*A List of Maximum-Period NLFSRs",* Cryptology ePrint Archive, Report 2012/166, March 2012, http://eprint.iacr.org/2012/166
4. Dubrova, E. "*A scalable method for constructing Galois NLFSRs with period* $2^n - 1$ *using cross-join pairs"*. Technical Report 2011/632, Cryptology ePrint Archive, November 2011. http://eprint.iacr.org/2011/632.
5. eSTREAM: eSTREAM – The ECRYPT Stream Cipher Project: http://www.ecrypt.eu.org/stream/
6. Golomb. "*Shift Register Sequences"*. Aegean Park Press (1982).
7. Massey, J.L. "*Shift-register synthesis and BCH decoding*". IEEE Transactions on Information Theory 15 (1969).
8. Maximov, A. and Biryukov, A. "*Two Trivial Attacks on Trivium*", Selected Areas in Cryptography, Lecture Notes in Computer Science, Vol.4876, Springer, 2007.
9. McDonald, C. and Pieprzyk, C. "*Attacking Bivium with MiniSat*", Cryptology ePrint Archive,Report 2007/040 (2007).
10. Raddum, H."*Cryptanalytic Results on Trivium*", eSTREAM, ECRYPT Stream Cipher Project, Report 2006/039 (2006).
11. Yun Tian, Gongliang Chen, Jianhua Li: "*On the Design of Trivium*". IACR Cryptology ePrint Archive (2009).

Usability Support Security Patterns

Susana Romaniz[1], Marta Castellaro[1], Juan Carlos Ramos[1], Ignacio Ramos[1]

[1] Facultad Regional Santa Fe - Universidad Tecnológica Nacional,
Lavaise 610 (S3004EWB) Santa Fe Argentina

{sromaniz, mcastell, jramos, iramos}@frsf.utn.edu.ar

Abstract. The main feature of secure software lies in the nature of processes and practices used to specify, design, develop and implement software. Security patterns applied the concept of pattern in the security realm. Its description helps to capture immediately the essence: what is the problem to which attends and what the proposed solution is. The different formats that exist for its description and the multiplicity of sources make its discovery demand effort that discourages the systematic use by potential recipients. This paper presents the prototype of a catalogue that seeks to establish a bridge between the knowledge and experience security experts and the needs of knowledge of software development teams.

Keywords: Security patterns. Security intelligence. Security patterns catalogue.

1 Security in the software development process

The main feature of secure software lies in the nature of the processes and practices used to specify, design, develop and deploy the software [1]. A project that adopts an improved security software development process incorporates a set of practices that reduce the number of exploitable flaws and bugs. Over time, these practices become more systematic, so it should decrease the likelihood that such vulnerabilities are present in the software at the moment that releases. The results in the field of research and experiences in the industry indicate the importance of reducing such potential vulnerabilities as early as possible within the software development lifecycle. The adoption of improved security processes and practices is much more profitable than the solution widespread today to develop and release patches for the operating software [2].

This early attention of the security has to do with the adoption of a set of activities that make possible the security integration in the software development lifecycle [3], including [4]: 1) identify security objectives, 2) apply security design guidelines, 3) create threat models, 4) conduct security architecture and design reviews, 5) complete implementation security reviews, and 6) run deployment security reviews.

We note the active development of tools and methods for testing that allows assessing the robustness and resilience of software products and their underlying infrastructure under attack conditions (for example, those used in penetration testing). Namely, there is intelligence domain of attacks that exploit vulnerabilities and compromise the security of software-intensive systems. All this produces a complex deal scenario.

There are criteria that help to secure software production [5]:

- Keep in mind that the security and cost of production of a software system depends strongly on the knowledge about its requirements [6].
- Include security treatment in each of the different stages of the software development cycle is an accepted criterion for improving the security of the final product. [7]
- Incorporate security patterns, which represent the best practices achieved by the industry in order to stop or limit security attacks [8].

In the secure software development process' case, security patterns are a way to bridge the gap between theory and practice. Although there are theoretical approaches, they are limited to relatively complex systems, and require a grade of knowledge and experience that is not available at the necessary level.

Then, we may infer that promote the use of security patterns to guide and drive the building of secure software development models, from the early stages of the process, is an approach that will allow us to ensure greater security in the behavior of a software product.

2 Security Patterns

The concept of "pattern" is known by the community as a solution to common problems in software development, whose effectiveness has been verified by solving similar problems in the past and that is reusable (applicable to different design problems in various circumstances). In the case of the security software aspects, which are found throughout all phases of development, its original definition [9] states that: "*Each pattern is a three-part rule, expressed as a relation between a certain context, a certain system of forces that occur repeatedly in this context, and a certain software settings that allows these forces to resolve themselves.*"

Extending the concept to software security, security patterns documented well known solutions to recurring problems of information security, allowing an efficient transfer of experience and knowledge. They apply the pattern concept to the realm of security, describing a particular recurrent security problem occurring in a specific context and presenting a well-proven generic solution accepted by the community of experts [10]. To make explicit the assumptions under which apply their solutions are applicable, reduce the risk of inappropriate use.

The solution proposed by a security pattern consists of a set of interacting roles that can be organized in multiple structures (applicable to the phases of requirements analysis, design, coding, testing or implementation, as appropriate pattern) concrete, as well as a process to create a particular structure in these [8]. According to what is expressed in [11] "a pattern defines a process and a thing: the 'thing' is created by the 'process'." Security patterns can be categorized according to a point of view associated with a software development lifecycle [12]. In this way there are patterns to the requirement phase, patterns to the design phase and patterns for the implementation phase. In general terms, guide the analysis through the context of security patterns allows integrate the problem addressed and the forces present that determine the problem.

Specifically, highlights the following advantages in the use of an approach to the use of patterns in security treatment: (i) express the basics security in a structured and understandable way; (ii) its representation is familiar to software developers and system

engineers, a key part of their audience; (iii) the patterns already are used to capture the knowledge about the system and the organization, then using the patterns to capture security knowledge help to improve security in the systems lifecycles, where results clearly needed. In particular regard to its development, in the last years different security groups have been specifying different patterns, as well as have made efforts for classification purposes [13, 14].

2.1 Security patterns description

Often refers to the model POSA (*Pattern-Oriented Software Architecture*) model developed by Buchman and others [15] to describe the context and the use of security patterns. But you can also see the existence of different descriptive models [16]; even, in many cases they are described so that the model is no strictly respected. These are discussed in detail in [17]. With regard on the format adopted for the description of security patterns, in general, it adopts the definition of a template, which contains a series of named elements and a defined scope. A good description will help to capture immediately the essence of a pattern, i.e., what is the problem to which attends and what is the proposed solution. It also offers all the details necessary to implement it and consider the consequences of its application.

It is important to note that not all fields are required. For example, in the case of some patterns is difficult or unnecessary to provide detailed descriptions of their structure, behavior and the implementation, because the information can be well integrated in the description of the solution. Likewise, in [18, 19] have defined other formats for the description of security patterns, also based on templates defined as sets of named elements and associated scope. A preliminary analysis of the different types of templates shows that there is not an exact match between the elements and scope, although all of them captured approximately the same aspects of security pattern to describe it. In [17] presented the results obtained with respect the variability of the aspects used for the formal definition of security patterns, which were obtained from a in-depth review of 364 security patterns collected from different sources, which were published during the years 1997-2012. These review also allowed us to see very high heterogeneity on the way of naming the aspects and there are no criteria of correspondence between the different names, such as exits between the descriptions of GoF and POSA [20].

2.2 Corporate and community knowledge compilation

But already there is no denying the need to address the problem of security products and services based on software, responsible for projects still face serious difficulties in addressing effectively the priorities identified in terms of security. What is the cause of this discrepancy? We cannot say that only due to ignorance about the existence of attackers who manage to exploit vulnerabilities in software. In fact, we can observe a significant disconnect between security experts and software development teams; the first are focused on the security of a system, while the seconds in building a system. For the latter, security is one of the non-functional goals with relevant, but just one of many.

Security patterns constitute a technology adopted for the description, communication and knowledge sharing, and proposed as a bridge that seeks to reduce this gap by capturing security expertise in the form of verified solutions to recurring problems. It is expected that they will be understood and used by the different members of development teams that are not security experts. Since the emphasis is on security, these patterns capture the strengths and weaknesses of different approaches in order to allow development teams to make informed decisions that balance security with other objectives.

We must emphasize that in this paper we have adopted the intelligence concept to refer to the set of practices that give rise to a collection of corporate knowledge, which is used to carry out proactive software security activities across the organization.

We decided to work in the organization and accessibility of that intelligence, seeking to generate a proposal for cataloguing security patterns. So we have considered the possibility of having resources that allow you to access a repository where reference to the set of defined security patterns in a systematic way of special importance. In this way, we hope to put at the disposal of the different actors involved in the software development process (project leaders, architects, designers, QA managers, programmers, testers) this knowledge drawn from the real world on the basis of the experience of specialists in security.

2.3 Cataloguing security patterns

At present, a multiplicity of sources exists where there are available different groups of security patterns defined throughout the time. This does that its discovery demands a level of effort that, normally, discourages to whom they are destinated to make use of them in systematic form. A centralized catalogue is a tool that acts as a starting point for the search and identification of one or more solutions to a security problem that is intended to resolve, expressed through security patterns. In this way it seeks to establish a bridge between the intelligence developed by security experts and the needs of knowledge of the software development teams.

To do this, we define as main requirement that the information offered by the catalogue be appropriate to "find" the security pattern. This information is extracted or inferred from the description of the pattern itself, and includes extensions that facilitate the categorization of the pattern according to established criteria. In the current design of the catalogue we adopted two criteria: (1) the ***security attributes*** impacted by the problem described; (2) the ***software development process phase*** to which applies the pattern. In this way, we hope that the information related with security patterns presented via the catalogue helps to capture immediately the essential aspects of them.

3 Proposal for cataloguing and its contribution

Our problem then is to define a way of structuring and indexing a catalogue of security patterns, so it could result quite easy to find a pattern that proposes a solution to an identified security situation, and from there access to the complete original reference of

the security pattern description. We define the set of attributes shown in Table 1 as the common attributes of security patterns that allow finding their references.

Table 1. Attributes for describing security patterns.

Nombre	Name of the original definition security pattern.
Objetivo	Problem that the security pattern meets. It is a response to a security problema.
Clasificación	Based on the phase of the software development process in which normally the pattern applies: requirements, analysis, design, coding, testing, implementation.
Aspecto seguridad afectado	Confidentiality, integrity, availability, accountability, non-repudiation
Keyworkds	They serve as a complementary reference to the security pattern.
Referencias	Links towards the documents and/or web pages where one finds the detailed description of the pattern.

Except for the 'Nombre' attribute, for the remaining attributes is necessary to make an analysis of the security pattern description in its original source, extract the attributes concepts associated with the selected attributes and perform the cataloguing of the security pattern. With respect to this concepts extraction process during the review of a security pattern, we found a guide in [17]; this paper summarized the results about the quality of the information included in the aspects that describe a security pattern, as well as the frequency of use of these aspects used along 67 publications of security patterns. In this regard, we can highlight the following as a guide to the extraction process:

- *Solution* (present in 87% of publications), is used to describe what security aspects attends the pattern and how it could be implemented;
- *Problem* (present in 84% of publications), in many cases includes abstract or oversimplified description of the problem;
- *Related Patterns and Consequences* (present in 75% of publications), require a good understanding of other security patterns as potential impact in the field of security so that they can be used as elements of distinction.
- *Context* (present in 49% of publications) generally includes a brief description, which makes it difficult to extract sufficient knowledge or even an idea about what the security pattern;
- *Known Use* (present in 46% of publications), described in what real-life cases you can use the pattern and provides guidance on its application domain.

4 Cataloguing process and prototype

Here by means of an example, the description of the process of incorporation of a security pattern to our catalogue. The selected pattern is called *Authenticated Session* [21], which is summarized in Figure 1. This process consists in the reading and analysis of the description security pattern conducted by a security expert, and in obtaining the data shown in Table 2, which result from the following references (the latter depends on the particular format adopted by the authors for the description of the security pattern and the quality of associated information): i) *Objetivo*: it is inferred from the Abstract section of the description; ii) *Clasificación*: the security expert who performs the reading and analysis of the security pattern inferred that the

Authenticated Session
(a.k.a. Server-Side Cookies, Single Sign-On)

Abstract

An authenticated session allows a Web user to access multiple access-restricted pages on a Web site without having to re-authenticate on every page request. Most Web application development environments provide basic session mechanisms. This pattern incorporates user authentication into the basic session model.

Problem

HTTP is a stateless, transaction-oriented protocol. Every page request is a separate atomic transaction with the Web server. But most interesting Web applications require some sort of session model, in which multiple user page requests are combined into an interactive experience. As a result, most Web application environments offer basic session semantics built atop the HTTP protocol. And the protocol itself has evolved to provide mechanisms -such as basic authentication and cookies- that allow session models to operate correctly across different Web browsers. An obvious use for session semantics is to allow users to authenticate themselves once instead of every time they access a restricted page. However, great care must be taken when using session semantics in a trusted fashion. Most session mechanisms are perfectly adequate for tracking non-critical data and implementing innocuous transactions. In such cases, if an end user circumvents the session mechanism, no harm is caused. But it is easy to make mistakes when applying session mechanisms to situations where accountability, integrity, and privacy are critical.

Solution

An authenticated session keeps track of a user's authenticated identity through the duration of a Web session. It allows a Web user to access multiple protected pages on the Web site without having to re-authenticate him/her-self on every page request. It keeps track of the last page access time and causes the session to expire after a predetermined period of inactivity. …..

Examples

Many significant Web banking and e-commerce applications rely on this pattern. Any site that enforces user authentication and does not store that information on the client uses something similar. ….

Related Patterns

· *Network Address Blacklist* – a related pattern that demonstrates a procedure for blocking a network address from further access attempts if a session identifier guessing attack is conducted.

· *Password Authentication* – a related pattern that presents the secure management of passwords, which are almost always used as the authentication mechanism for this pattern.

References

[1] Coggeshall, J. "Session Authentication". http://www.zend.com/zend/spotlight/sessionauth7may.php, May 2001.

[2] Cunningham, C. "Session Management and Authentication with PHPLIB". http://www.phpbuilder.com/columns/chad19990414.php3?page=2, April 1999.

[3] Kärkkäinen, S. "Session Management". *Unix Web Application Architectures*. http://webapparch.sourceforge.net/#23, October 2000.

Figure 1. Selected security pattern to cataloguing: *Authenticated Session* [21].

Table 2. Data for cataloguing the security pattern selected.

Nombre	**Authenticated Session**
Objetivo	Allow the access to multiple pages with restricted access, without having to re-authenticate again and again. Keep authentication information in the system's navigation
Clasificación	Design
Aspecto de seguridad afectado	Accountability, Integrity.
Keywords	Authentication, Single Sign-On
Referencias	• Security Patterns Repository v1.0.pdf • Cunningham, C. "Session Management and Authentication with PHPLIB". http://www.phpbuilder.com/columns/chad19990414.php3, (Rev. Mayo 2013). • Kärkkäinen, S. "Session Management". *Unix Web Application Architectures*. http://webapparch.sourceforge.net/#23, October 2000. (Rev. Mayo 2013)

same applies to the design phase, in the aspect of 'user authentication'; iii) *Aspecto de seguridad afectado*: in the final paragraph of section description Problem refers to "But it is easy to make mistakes when applying session mechanisms to situations where accountability, integrity, and privacy are critical", accordingly, the security pattern seeks to address these shortcomings; iv) *Keywords*: inferred from Name and Know-As-As sections; v) *Referencias*: from the references listed in the References section, selecting those sources whose availability at the time of cataloguing can be verified. Obtained these data, we proceed to cataloguing the security pattern selected for example, using the prototype that we describe in the following section.

To realize this proposal, we analyzed extended use alternative tools that allow us to its implementation. We define build a prototype using a tool for management of bibliographic references: 'JabRef'. This is a very low-cost alternative to test the idea which, if it actually works, can then be extended to massive tools, as for example a web application with features 'wiki'.

JabRef is configurable bibliographic reference management software. It use native format BibTex (a text-based and independent of the style file format) to define lists of bibliographic items, articles, thesis, etc. For the generation of the proposed catalogue we take advantage of this facility for recording references in this manager; in addition, the existence of distinct and specific attributes in each pattern allows us to generate a special entry type (Security Pattern type) with their respective fields and general and/or optional information. Another important feature of JabRef is the use of LaTeX, which allows us to transfer automatically and without major difficulties (given the use of well-defined parameters) the current catalogue to any other type of software that allows the entry of the same language: databases, Wikipedia pages, another specific or general purpose manager, etc. Figure 2 shows a screen from the JabRef interface, where:

- *Groups* (section 1). Created groups that correspond to the phases of software development in which the pattern is normally applicable. Selecting a group in section 2 (Entry) will appear the patterns that correspond to that classification.
- *Entry* (section 2). Entries loaded at the base are presented as a table with the fields as columns, which can be used to select a criterion of order by pressing the desired column. You can see all the references or related files by right clicking

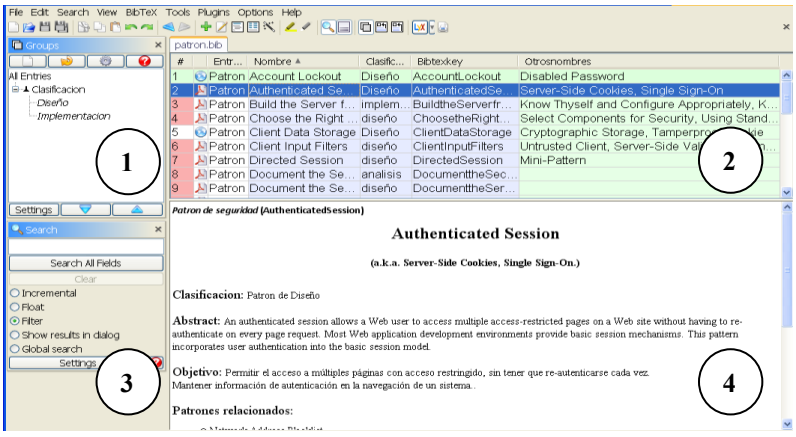


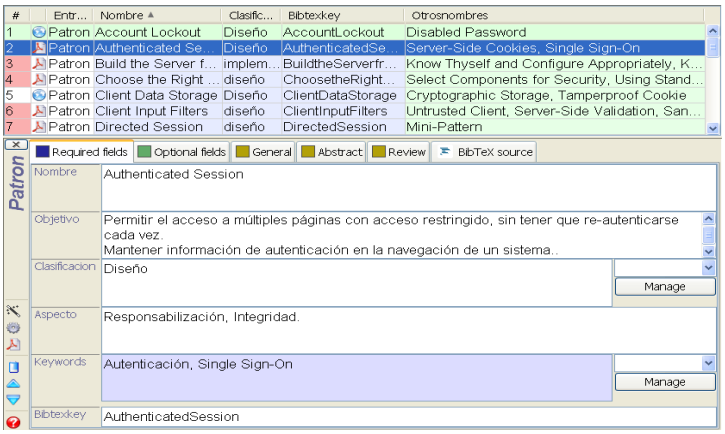
Figure 2. JabRef Catalogue Interface.

on the desired row in the column '*File*' (second column), folding a list of values. If you select one of these, will go to the required reference, whether a document available locally or on the Internet.

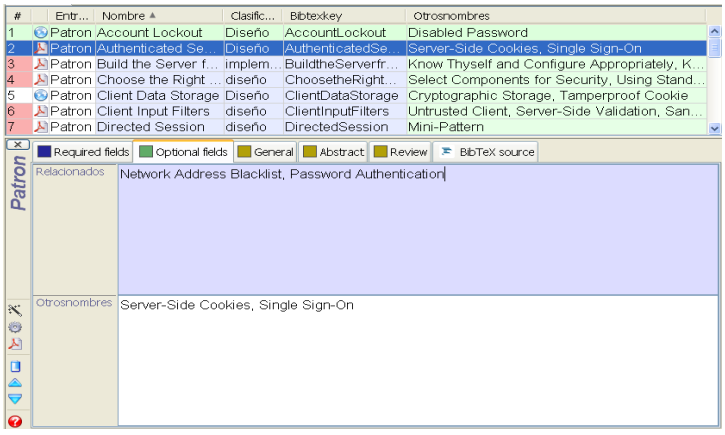
- *Search* (section 3). It allows to search according to an attribute and filters inside the entire catalogue. After a search and already showing the selected patterns that contain the phrase or word searched, when you double-click on them and navigating through the different tabs of the query, words appear highlighted in blue.
- *Preview* (Section 4). It shows the information of the security pattern selected in PDF format, which is a friendly representation of the contents of the pattern by default. By pressing the right mouse button you can print preview.

This figure shows the way how catalogued information relating to the Authenticated Session pattern [21] is displayed, where there are the attributes proposed to categorize a security pattern.

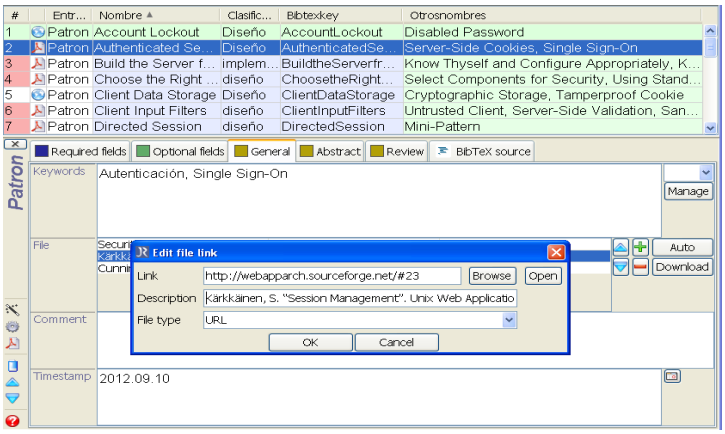
Cataloguing a security pattern: Figure 3 shows some of the facilities offered by the developed prototype to incorporate information from the attributes listed in Table 2: Figure 3.a): Loading the 'Required Fields' (mandatory): name, objective, classification, security issues, key words and Bibtexkey (a peculiarity of JabRef and



a) Required Fields.



b) Optional Fields.



c) Loading 'Keywords' and 'Referencias'.

Figure 3 Interfaces for loading data for cataloguing attributes *Authenticated Session* security pattern.

bibtex references, which is used for references to the security pattern); Figure 3b) Information is supplemented with related patterns and other well-known names (in both cases, if any); Figure 3.c): Defining 'Keywords' and 'Referencias' for the security pattern, taking advantage of the JabRef's facility for linking local and external files as 'Files'.

Searching a security pattern in the catalogue: Figure 4 shows the result of a search for references to security patterns applicable to the design phase and that attend to the confidentiality as a security attribute.

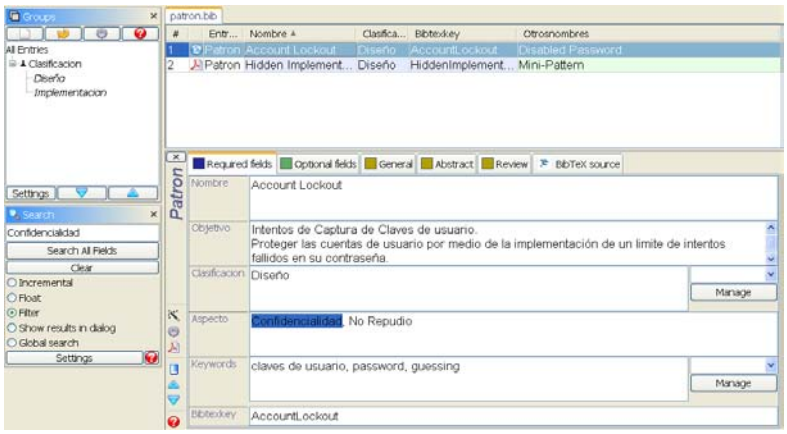


Figure. 4. Results of a search in the security pattern catalogue.

5 Conclusions and Future Work

This proposal aims to provide to the security community these quick references, and go replenishing it as it's used. The tool chosen to build a catalogue of references

helps us to test in a simple way, and eventually modify, the proposed structure and criteria, and consider the possibility of replace with another with higher performance and style. This prototype allows us to define the bases to make an extension to a wiki-style web tool, which will be available to the whole community, and also to be updated by it. In addition to the web tool, it will be necessary to propose and publish criteria to incorporate new references to our catalogue, so that all entries follow these agreed common criteria. This is the work that we intend to address in a next phase.

References

1. McGraw, G., "Software Security: Building Security In." Addison-Wedsley. EEUU. (2006).
2. Romaniz, S., "Buenas prácticas de elicitación de los requerimientos de seguridad". IV Congreso Iberoamericano de Seguridad Informática -CIBSI2007-, Argentina (2007).
3. Meier, J. et al., "Security engineering explained." (2005). Available in http://www.microsoft.com/download/en/confirmation.aspx?id=20528.
4. Castellaro, M. y otros, "Hacia la Ingeniería de Software Seguro." XV Congreso Argentino de Ciencias de la Computación, CACIC2009. Argentina. (2009).
5. Solinas, M., "Elicitación y trazabilidad de requerimientos utilizando patrones de seguridad." Universidad Nacional de La Plata. Argentina. (2012). Available in http://sedici.unlp.edu.ar/bitstream/handle/10915/421/Documento_completo.pdf?sequence=1.
6. Garvin, D., "What does product quality really mean?" Sloan Management Review, Vol 26, No 1. (1984).
7. Romaniz, S. y otros, "La seguridad como aspecto organizacional y transversal en proyectos de Sistemas de Información." 38 Jornadas Argentinas de Informática 38JAIIO. Argentina. (2009).
8. Schumacher, M. et al., "Security Patterns: Integrating Security and Systems Engineering." John Willey & Sons Inc. EEEUU. (2006).
9. Coplien, J.: "Design Pattern Definition - Software Patterns." Available in http://www.hillside.net/component/content/article/50-patterns/222-design-pattern-definition.
10. Schumacher, M.: "Security engineering with patterns-origins, theoretical model, and new applications." Springer-Verlag. (2003).
11. Alexander, C.: "The Timeless Way of Building." Oxford University Press. EEUU. (1979).
12. Yoshioka, N. et al.: "A survey on security patterns." Progress in Informatics. (2008).
13. Fernandez, E. et al. "Classifying Security Patterns." Progress in WWW Research and Development Volume 4976. (2008).
14. Washizaki, H. et al. "Improving the Classification of Security Patterns." 20th International Workshop on Database and Expert Systems Application, DEXA'09. Austria. (2009).
15. Buschmann, F. et al. "Pattern-Oriented Software Architecture: A System of Patterns." Chichester, UK: Wiley, 1996.
16. Schumacher, M. et al., "Security engineering with patterns." Proceedings of the Conference on Pattern Languages of Programs, pp. 1–17. (2001).
17. Bunke, M. et al. "Organizating Security Patterns Related to Security and Pattern Recognition Requirements." International Journal on Advances in Security, vol 5 no 1 & 2 (2012).
18. Kienzle, D.: "Security Patterns Template and Tutorial version 1.0." (2008). Available in http://www.scrypt.net/~celer/securitypatterns/template%20and%20tutorial.pdf
19. Blakley, B. et al. "Security Design Patterns." The Open Group. (2004). Available in https://www2.opengroup.org/ogsys/catalog/g031.
20. Henninger, V. et al. "Software pattern communities: Current practices and challenges." Proceedings of the Conference on Pattern Languages of Programs PLOP. New York, NY, USA. ACM, 2007, pp. 14:1–14:19.
21. Darrell M. Kienzle et al. "Security Patterns Repository Version 1.0." Available in http://www.scrypt.net/~celer/securitypatterns/repository.pdf

Analizador de Intents en Android

Joaquín Erario, Christian Rovera, Francisco Bavera

Departamento de Computación
Universidad Nacional de Rí o Cuarto
Río Cuarto, Argentina
pancho@dc.exa.unrc.edu.ar

Resumen Existen numerosos reportes de vulnerabilidades detectadas en el sistema operativo Android. Si bien muchas de estas vulnerabilidades fueron solucionadas rápidamente, se detectó una, que hasta la fecha, no ha sido solucionada: vulnerabilidades por el uso de "Intent" explícitos. Un "Intent" explícito es un método que puede ser utilizado en aplicaciones Android para invocar desde una aplicación a otra aplicación o servicio determinado explícitamente. Esta invocación puede incluir el paso de algún tipo de información (posiblemente sensible o confidencial). Esta forma de invocar Intents puede ocasionar una vulnerabilidad, ya que, la aplicación o servicio que se invoca puede ser modificada (de forma maliciosa o no) y esta modificación puede permitir manipular de manera indeseada la información recibida. Por ejemplo, una aplicación que envía un intent que agregue un contacto de la agenda o un mensaje para publicar en una red social y la aplicación de destino no es la esperada se puede filtrar información sensible o confidencial sin el consentimiento del usuario. En este trabajo se presenta una herramienta para garantizar que esta vulnerabilidad no pueda ser explotada. El enfoque utilizado segue los lineamientos de la técnica conocida como integridad de control de flujo.

Palabras Clave: Integridad de Control de Flujo, Seguridad, Verificación, Lenguajes, Programas.

1. Introducción

En los últimos años, el sistema operativo Android, inicialmente pensado para teléfonos móviles, fue creciendo cada vez más en cuanto a popularidad debido a una de sus mejores características: la libertad.
Es que Android es un sistema operativo completamente libre. Es decir, no hay que pagar absolutamente nada ni para programar en el ni para poder instalarlo en un teléfono. Lo que lo hace muy popular entre desarrolladores, que deben pagar muy poco para lanzar una aplicación, y fabricantes de celulares, que ahorran a la hora de elegir el sistema operativo para el teléfono que quieren lanzar al mercado.

Otra de las ventajas de ser un sistema operativo libre es que cualquier persona puede descargar el código fuente, inspeccionarlo, modificarlo y finalmente compilarlo. Por lo que, como cualquier software libre, es más fácil detectar errores rápidamente y por ende solucionarlos, esto favorece mucho a la seguridad en el sistema operativo ya que las vulnerabilidades que van surgiendo se van reparando constantemente. Sin embargo, hay ciertos aspectos en cuanto a seguridad, que a pesar de esto, aún no han sido cubiertos. Y esta fue nuestra motivación para llevar a cabo este trabajo: Poder solucionar alguna de estas fallas de seguridad a nivel aplicación que a la fecha aún no han sido solucionadas.

Existen reportes de numerosas cantidades de vulnerabilidades detectadas en el sistema operativo Android, pero como se mencionó anteriormente, al ser un sistema de código libre, estas vulnerabilidades son solucionadas rapidamente luego de ser reportadas. Aunque se encontró una que hasta la fecha, no ha sido solucionada: Vulnerabilidades al usar "Intent" explícitos.

Brevemente, un "Intent" es un método que pueden ser utilizados en aplicaciones Android para invocar desde una aplicación a otra. Esta invocación puede incluir el paso de algún tipo de información y puede ser de manera implícita o explícita. En la forma implícita,el programador no indica que aplicación en concreto quiere invocar sino que simplemente indica el tipo de información que quiere que sea procesada y el sistema operativo se encarga de elegir las aplicaciones correctas para que luego el usuario opte por la que desee. Por otro lado, en la forma explicita, el programador indica específicamente que aplicación invocar.

Esta segunda forma de hacer el Intent puede ocasionar una vulnerabilidad, ya que, la aplicación que se invoca (aplicación destino) puede ser modificada (de forma maliciosa o no) y esta modificación puede permitir manipular de manera indeseada la información recibida. Por ejemplo, si tenemos una aplicación que envía un intent que agregue un contacto de la agenda o un mensaje para publicar en una red social y la aplicación de destino no es la esperada se puede filtrar información confidencial.

A continucación se presenta una breve introducción a Android, seguido de una descripción de algunas vulnerabilidades reportadas. Luego, se describe brevemente la técnica que motivó este trabajo: Integridad de Control de Flujo. Seguido de la presentación de la herramienta. Para finalizar se presentan las conclusiones.

2. Android

Android es un sistema operativo basado en Linux, diseñado principalmente para dispositivos móviles con pantalla táctil como por ejemplo tablets y smarthphones. Fue desarrollado por Android Inc. con el respaldo económico de Google que en el 2005 termina comprando a la empresa. Finalmente, en octubre del 2008 se vendió por primera vez un celular con este sistema operativo.

2.1. Arquitectura de Android

Los componentes principales de Android son cinco:

- Aplicaciones: Este primer componente o nivel, esta compuesto por el conjunto de todas las aplicaciones instaladas en una máquina Android y deben correr en la máquina virtual Dalvik para "garantizar" la seguridad del sistema. Generalmente las aplicaciones Android están escritas en Java aunque también se pueden programar en C++ utilizando el kit de desarrollo Android NDK (Native Development Kit).
- Marco de trabajo de aplicaciones: Los desarrolladores tienen acceso a los mismos APIs del framework que usan las aplicaciones base (sensores, barra de notificaciones, servicios, etc...). Esta capa está diseñada para simplificar la reutilización de componentes. Cualquier aplicación puede publicar sus capacidades y cualquier otra aplicación puede luego hacer uso de estas (respetando siempre las reglas de seguridad del framework). Este mismo mecanismo permite que los componentes sean reemplazados por el usuario.
- Bibliotecas nativas: Android incluye además un conjunto de librerías de C/C++ que son usadas por varios componentes del sistema. Y son expuestas también a los desarrolladores por medio del marco de trabajo de aplicaciones. Se pueden destacar bibliotecas como: System C library, Media Framework, etc..
- Runtime de Android (Máquina Virtual Dalvik): Debido a las limitaciones de memoria y procesador de los dispositivos móviles donde ha de correr Android no fue posible utilizar la máquina virtual estándar de Java para la ejecución de aplicaciones, Google debió crear una nueva máquina virtual Dalvik que funcione mejor ante estas limitaciones. Algunas características de la máquina virtual Dalvik que ayudan a optimizar recursos son:
  - Ejecuta ficheros Dalvik ejecutables (ficheros con extensión .dex). Este formato está optimizador para ahorrar memoria.

- Basado en registros, en lugar de estar basada en una pila.
- Cada aplicación corre en su propio proceso Linux con su propia instancia de la máquina.
- Delega al kernel de Linux algunas funciones como threading y el manejo de la memoria a bajo nivel.

- Núcleo Linux: El núcleo de Android está formado por el sistema operativo Linux. Esta capa proporciona servicios como la seguridad, el manejo de la memoria, el multiproceso, la pila de protocolos y el soporte de drivers para dispositivos. Es la única capa dependiente del hardware ya que actúa como capa de abstracción entre el hardware y la pila de protocolos.

2.2. Aplicaciones

Las aplicaciones de Android, como ya mencionamos, se pueden programar tanto en C++ como en Java. Cuando se compilan los programas, el Kit de Desarrollo nos arma un archivo .APK. Los programas compilados son programas en Bytecode para la máquina virtual Dalvik.
Los archivos .APK son los que nos permiten instalar una aplicación en el celular. Contienen una serie de instrucciones a ejecutar y además otros recursos como imágenes, sonidos y archivos .xml cómo por ejemplo el AndroidManifest.xml.
Cada APK se asocia a un proceso único, que proporciona el ambiente de ejecución de los componentes. De los cuales, uno es el componente inicial del programa.
Cuando se ejecuta una aplicación se le asigna un proceso Linux y un único hilo de ejecución (thread), así todos sus componentes corren sobre el mismo proceso y thread.

2.3. Seguridad en Android

La seguridad es un aspecto clave en cualquier sistema. Si nos descargamos una aplicación maliciosa a nuestro celular, esta podría robarnos contactos, enviar sms por su propia cuenta y sin nuestra autorización o hasta incluso saber donde estamos posicionados físicamente utilizando datos del gps del sistema.
Android propone un esquema de seguridad para proteger a los usuarios, sin tener que imponer un sistema centralizado en alguna empresa (como si lo hace el sistema operativo de iPhone por ejemplo). Este esquema esta basado en tres pilares fundamentales: la seguridad de Linux, la firma digital de la aplicación y un modelo de permisos de acceso a partes del

sistema. Este último componente establece que muchas decisiones importantes relativas a la seguridad recaigan en el usuario final.

Seguridad Linux Como se comentó en la sección anterior, Android está basado en Linux, por lo tanto, se aprovecha la seguridad que incorpora este sistema operativo. De esta manera Android puede impedir que las aplicaciones tengan acceso directo al hardware o interfieran con recursos de otras aplicaciones.

Firma digital de las aplicaciones Las aplicaciones deben ser firmadas con un certificado digital que identifique al autor. Esto nos permite ver que aplicaciones se supone que sean más confiables que otras. Además, la firma digital nos garantiza que los archivos de una aplicación no han sido modificados. Si se opta por modificar la aplicación, está deberá ser firmada nuevamente. Generalmente este certificado digital no es firmado por alguna autoridad de certificación.

Modelo de permisos: Android Manifest Si queremos que una aplicación tenga acceso a partes del sistema que pueden comprometer la seguridad del sistema necesitamos utilizar un modelo de permisos, de forma el usuario puede conocer los riesgos antes de instalar la aplicación. Para lograr esto, se utiliza el archivo Android Manifest.
Cada aplicación de Android debe contener un archivo AndroidManifest.xml (con exactamente este nombre) en su directorio raíz. Este archivo le presenta información esencial sobre la aplicación al sistema operativo, información que el sistema debe tener antes de poder correr el código de la misma.Entre otras cosas, el AndroidManifest.xml tiene:

- El nombre del paquete Java que sirve como identificador único para las aplicaciones del sistema operativo.
- Componentes de la aplicación.
- Permisos que la aplicación va a solicitar al momento de su instalación. Pueden ser tanto como para acceder a la API o interactuar con otras aplicaciones.
- Bibliotecas con las que debe linkear.
- Nivel mínimo de la API de Android requerido para su funcionamiento.

2.4. Vulnerabilidad por Intent Explícitos

La interacción entre componentes en un sistema Android están dados por pasaje de mensajes llamados Intents. Estos mensajes están formados

por una dirección o nombre de destino e información relacionada con la acción a ejecutar.
Cuando un componente envía un Intent, el receptor del mismo iniciará una actividad, mensaje broadcast o servicio que ejecutarán una acción.
Hay dos tipos de intents posibles, por un lado tenemos los intents implícitos donde no se indica el destino del mensaje, sino que solo se indica el tipo de mensaje (texto plano por ejemplo) y entonces todas las aplicaciones que declaren en su AndroidManifest que pueden recibir ese tipo de mensajes, aparecerán en una lista para que el usuario decida que aplicación es la que recibirá el mensaje. Un claro ejemplo de esto, es cuando tenemos más de un navegador web instalado en el celular y hacemos click en algún link que nos lleve a una dirección web. En ese momento, nos aparecerá una lista con todos los navegadores instalados de la cual debemos elegir cuál queremos usar.
Por otro lado, tenemos los intents explícitos, aquellos a los que si queremos indicarle específicamente a quién va dirigido el mensaje, es decir, el nombre del paquete de la aplicación receptora.
La posible vulnerabilidad de este tipo de intents, radica en que al indicar solo el nombre del paquete de la aplicación, es posible cambiar la aplicación receptora por otra con el mismo nombre y cuyo comportamiento no es el que nosotros deseamos.

Para entender mejor esta vulnerabilidad veamos un simple ejemplo: dada una agenda de contactos que posee la funcionalidad de enviarle los contactos favoritos a otra aplicación de mensajería (Similar a aplicaciones como Viber o Whatsapp). Esta aplicación de mensajería, recibe los contactos de esta nueva agenda y los utiliza para poder enviarle y recibir mensajes de ellos.El problema esta en qué si se cambia esta aplicación de mensajería, puede ocurrir que en lugar de recibir los contactos y guardarlos para luego poder comunicarse con ellos, los envié a una dirección de correo y así nos robe los contactos de la agenda.

3. Integridad del control de flujo

Las computadoras con frecuencia son objeto de ataques externos que apuntan a controlar el comportamiento de algún software. Generalmente estos ataques llegan como datos a través de un canal de comunicación normal y, una vez que residen en la memoria del programa, explotan defectos preexistentes en el software. Explotando dichos defectos, el atacante desestabiliza y toma control del comportamiento del software. Por ejemplo, un desbordamiento del buffer en una aplicación puede resultar en una

llamada a una función sensible del software, posiblemente una función que la aplicación no fue diseñada para su uso.
Las políticas de la integridad de control de flujo [9,10] dictan que la ejecución del software deben seguir un camino de su grafo de control de flujo creado de antemano. El grafo en cuestión puede ser definido por análisis de código, análisis de los ejecutables o mediante perfiles de ejecución.
Hay varias formas de llevar dicho control, pero en el presente trabajo nos centraremos en la instrumentación de código debido a que es el método utilizado en nuestra herramienta.

3.1. Instrumentación de código

La instrumentación de código es la inserción de código con el fin de asegurar la perfomance de un sistema, diagnosticar errores y escribir información del flujo del programa. La instrumentación le otorga a un programa la de incorporar:

- Seguimiento de código: recibiendo mensajes informativos acerca de la ejecución de una aplicación en tiempo de ejecución.
- Depuración y un estructurado manejo de excepciones: búsqueda y corrección de errores de programación en una aplicación bajo desarrollo.
- Profiling: un medio por el cual los comportamientos dinámicos de los programas pueden ser medidos durante un ejecución de prueba con una entrada representativa. Esto es útil para probar propiedades de un programa que no puede ser analizadas estáticamente con suficiente precisión, como por ejemplo análisis de aliasing.
- Contadores de rendimiento: componentes que permiten el seguimiento de la performance de una aplicación.
- Registro de datos: componentes que permiten el registro y seguimiento de la mayoría de los eventos en la ejecución de la aplicación.

4. La Herramienta

Teniendo en cuenta la vulnerabilidad explicada en la sección anterior sobre los intents explícitos, se decidió crear una herramienta que pudiera de alguna forma solucionar este posible problema en las aplicaciones. La herramienta tiene la funcionalidad de controlar que el flujo de la información mediante intents sea el requerido por el usuario. Para eso se introduce en la aplicación que envía la información, antes de cada invocación de intent, una verificación para garantizar que efectivamente el mensaje sea atrapado por la aplicación deseada.

Para lograr esto se trabajó sobre el control de flujo de la aplicación, más precisamente se instrumentó código a la aplicación que envía el intent. A grandes rasgos, este código es una función que toma como parámetro el nombre de una aplicación y un identificador único (id) de la misma. Este id es obtenida calculando un código hash para luego checkear en una lista confiable presente en el celular si el id de esa aplicación es el mismo que el id actual de la aplicación con ese nombre. De esta forma se detecta si el receptor del mensaje fue modificado o no.

La herramienta y su documentación esta disponible en `https://sourceforge.net/projects/`

4.1. El Proceso

A continuación se detalla el proceso realizado por la herramienta para generar las verificaciones. Este proceso se realiza cuando la aplicación es instalada en el dispositivo Android.

Conversión a código Jasmin En primer lugar, cuando el usuario elige el programa que desea verificar que sus intents sean correctos, la herramienta transforma el archivo .APK (el programa a instalar en el dispositivo) a código Jasmin [6] (una representación intermedia de código bytecode), para eso invoca una serie de scripts (basados en la herramienta Dex2Jar [3]:

- **d2j-dex2jar.sh**: Extrae el classes.dex del apk y lo convierte en un archivo .jar.
- **d2j-asm-verify.sh**: Verifica que el .jar creado sea correcto.
- **d2j-jar2jasmin.sh**: Transforma el código .jar en código Jasmin.

Obtención de la Id de la aplicación Para poder controlar que la aplicación receptora del mensaje o intent no sea modificada es necesario asignarle una id única. Para esto se creó un script en Python que obtiene el código hash de un archivo, de esta forma se le obtiene el hash al archivo instalador (.apk) de la aplicación receptora que si recibe un mínimo cambio entonces su id cambiará. Para obtener el código hash del .apk se utiliza la librería hashlib de Python.

Inserción de las Verificaciones Dinámicas Se introducen las verificaciones dinámicas que chequean que el id de cada aplicación en tiempo de instalación se correspondan con el id de la aplicación en tiempo de ejecución. Estas verificaciones se introducen en el código Jasmin.

Para esto se implementó un método Check en Jasmin que toma como parámetros el nombre de la aplicación que se va a checkear y el id o código hash del mismo que espera. Luego busca en un archivo auxiliar instalado en el dispositivo donde se encuentran los nombres de las aplicación instaladas con nuestra herramienta y su código hash actual, que si el nombre de la aplicación pasada como parámetro coincide con alguno del archivo, sus ids también deben coincidir.

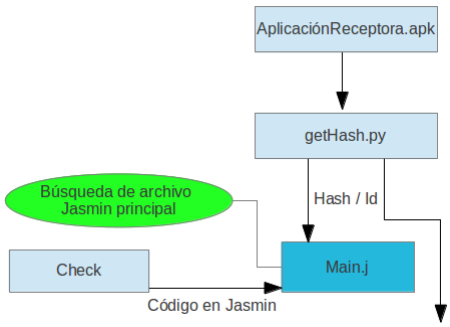


Figura 1. Instrumentación de código en el archivo principal

Finalmente se introduce antes de cada invocación a un intent que se desea controlar la llamada a esta función. En pseudocódigo, es una simple llamada a la función check: `check(nombrePaquete,IdEsperado)`.

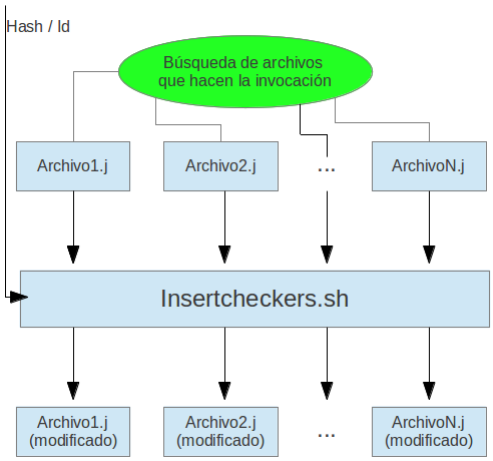


Figura 2. Script insertcheckers.sh: Instrumentación de código en el resto de los archivos que invocan a la aplicación receptora

Reconstrucción del .APK Una vez finalizada la inserción de código solo resta reconstruir el .APK para ello se utilizan algunos scripts de la herramienta Dex2Jar:

1. **d2j-jasmin2jar.sh**: Reconstruye el .jar.
2. **d2j-asm-verify.sh**: Nuevamente se verifica que el .jar sea correcto.
3. **d2j-jar2dex.sh**: Se reconvierte a código dex.

Posteriormente se crea una copia de seguridad del .apk y se le reemplaza el dex original con el modificado.
Por último, como Android necesita que los .apk estén firmados para que te los permita instalar, se vuelven a firmar.

La Verificación El proceso de verificación en tiempo de ejecución, es decir, la verificación cuando se ejecuta la aplicación es sencilla. El código instrumentado ejecuta un método qie verifica si el id actual del intent implícito que se ejecutará a continuación es igual al id en tiempo de instalación. En caso de no coincidir no se permite la ejecución del intent.

4.2. Ejemplos

En primer lugar se implementó una agenda de contactos donde se almacenan números telefónicos que pueden ser usadas por otra aplicaciones

para el envío de mensajes.
Si seleccionamos un contacto de la agenda, automáticamente este se lo envía a la aplicación de mensajería quién almacena el número y luego de que se escriba el mensaje, envía su contenido al destinatario.

Como segundo ejemplo se implementó una aplicación de logueo a homebanking de distintos bancos. La aplicación permite ingresar usuario y contraseña y elegir a que banco se desea acceder para luego enviar al modulo del banco seleccionado los datos de sesión y consultar saldo o movimientos. Este módulo podría ser modificado para que los datos del cliente se envíen ocultamente por email, o directamente que se hagan transferencias no deseadas.

Con estas dos aplicaciones se pudo apreciar que el enfoque es factible para reforzar la integridad de control de flujo en aplicaciones Android con la herramienta presentada. En abos casos se detectó cuando las aplicaciones fueron cambiadas y/o modificadas.

5. Conclusiones

El desarrollo de esta herramienta se puede fundamentar fácilmente revisando las vulnerabilidades que están presentes actualmente en el sistema operativo Android y en la presunción (fundamentada en la experiencia hasta el momento) de que se detectarán nuevas vulnerabilidades en el futuro. Por ejemplo, es posible escalar privilegios y modificar una aplicación, que se encuentre presente en el celular, receptora de algún intent o mensaje explícito. También puede ocurrir que la aplicación sea modificada por una actualización de la aplicación (una actualización que implique un cambio defuncionalidad no deseado).

La herramienta presentada permite controlar por fuera del sistema operativo y darle más garantía a los programadores de Android, de manera autómatica, verificando que los intent tengan el destinatario esperado. En otras palabras, la herramienta garantiza la integridad del control de flujo relacionada con los intents implicitos.

La solución presentada en el presente trabajo tiene la desventaja que no es una solución a nivel sistema operativo por lo que se necesita sí o sí de la herramienta realizada. Esta desventaja no es tan negativa por el hecho de que son escasas las veces que se requieren realizar este tipo de checkeos por lo que no es tan molesto tener que recurrir a la herramienta para lograrlo. Además necesita también de otra base confiable, en este caso de un instalador de aplicaciones alternativos que vaya actualizando y agregando los ids de las aplicaciones instaladas en un archivo

confiable. Por otro lado, la herramienta tiene la ventaja que se puede modificar fácilmente o incluso utilizarla para realizar otro tipo de chequeos o modificaciones de las aplicaciones.

5.1. Trabajos futuros

Los trabajos más relevantes que se deben realizar para mejorar la herramienta y extender su funcionalidad son:

- extender el método de verificación para incluir los intent implícitos. En Android cuando se hace un intent implícito se crea una lista de aplicaciones que pueden manejar el tipo de dato enviado (por ejemplo, texto plano) para que el usuario elija la aplicación que reciba el intent. Para garantizar la aplicación/es autorizadas a recibir determinado intent implícito es necesario poder generar la lista dinámicamente (y no tomar la lista generada autómaticamente por Android).
- La herramienta esta programada en bash de Linux y en Python ambos estan disponibles para Android, por lo que es factible hacerla nativa para Android. Sin embargo habría que modificar o sustituir el instalador de aplicaciones de Android por un instalador "seguro" para darle mayor grado de confiabilidad.
- Es necesario estudiar la posibilidad de implementar este enfoque para garantizar la integridad del control de flujo desde el receptor de los intent. Es decir, que el receptor sea el que verifique si el intent fue disparado por una aplicación autorizada.

Referencias

1. Web oficial de desarrollo para Android, *http://developer.android.com/index.html*
2. Android Argentina, *http://androidargentina.com.ar/*
3. Web oficial de Dex2jar, *http://code.google.com/p/dex2jar/*
4. ¿Qué es Android?, *http://www.xatakandroid.com/sistema-operativo/que-es-android.*
5. Wikipedia: Android, *http://es.wikipedia.org/wiki/Android*
6. Web de Jasmin, *http://jasmin.sourceforge.net/about.html*
7. Android Security, *https://source.android.com/tech/security/*
8. Introduction to Instrumentation and Tracing *http://msdn.microsoft.com/en-us/library/aa983649%28VS.71%29.aspx*
9. Martín Abadi; Mihai Budiu; Úlfar Erlingsson y Jay Ligatti. *Control-Flow Integrity Principles, Implementations, and Applications.* ACM Journal Vol V, Febrero 2007.
10. Chao Zhang; Tao Wei; Zhaofeng Chen; Lei Duan; László Szekeres; Stephen McCamant; Dawn Song y Wei Zou. *Practical Control Flow Integrity & Randomization for Binary Executables.*

II WORKSHOP DE INNOVACIÓN EN EDUCACIÓN EN INFORMÁTICA

- WIEI -

CACIC2013

II WORKSHOP DE INNOVACIÓN EN EDUCACIÓN EN INFORMÁTICA

- WIEI -

ID	Trabajo	Autores
5672	**Experiencia de utilización de Herramientas Colaborativas para la enseñanza y el aprendizaje de la Programación de Computadoras**	Edith Lovos (UNRN), Alejandro Gonzalez (UNLP), Rodolfo Bertone (UNLP)
5840	**Enseñanza de técnicas de elicitación de requerimientos**	Alejandro Oliveros (UADE), Javier Zuñiga (UADE), Sergio Corbo (UADE), Patricia Forradellas (UADE), Sandra Martínez (UADE)
5709	**Generando Entornos de Investigación y Desarrollo utilizando Redes Inalámbricas de Sensores (WSN)**	Eduardo Omar Sosa (UNaM), Diego Alberto Godoy (UGD), Edgardo Belloni (UGD)
5732	**El desarrollo de la comprensión lectora en las carreras de Informática**	Sonia V. Rueda (UNS)
5737	**Extensión del Lenguaje y Modelo Simplesem con Soporte para Paralelismo**	Lucas L. Diez de Medina (UNC), Gustavo Wolfmann (UNC), Orlando Micolini (UNC)
5740	**Conformando repositorios de datos de la comunidad educativa en la Universidad Nacional de La Plata Un caso de estudio**	Javier Diaz (UNLP), Maria Alejandra Osorio (UNLP), Ana Paola Amadeo (UNLP)
5748	**Expiriences with educational robotic**	Anibal Lopes Guedes (uffs), Fernanda Lopes Guedes (IF-SUL)
5687	**Desafíos y herramientas para la enseñanza temprana de Concurrencia y Paralelismo**	Laura De Giusti (UNLP), Fabiana Leibovich (UNLP), Mariano Sanchez (UNLP), Franco Chichizola (UNLP), Marcelo Naiouf (UNLP), Armando E. De Giusti (UNLP)

II WORKSHOP DE INNOVACIÓN EN EDUCACIÓN EN INFORMÁTICA

- WIEI -

ID	Trabajo	Autores
5838	**Una propuesta para la incorporación de Cloud Computing a la currícula de Grado**	Nelson R. Rodriguez (UNSJ), María Antonia Murazzo (UNSJ), Daniela Villafañe (UNSJ), Francisca Adriana Valenzuela (UNSJ), Adriana Martin (UNSJ), Susana Beatriz Chavez (UNSJ)
5782	**Propuesta de una metodología para una rápida enseñanza de circuitos lógicos y de su integración en una UCP en carreras de Informática**	Mario Carlos Ginzburg (UAI)
5857	**FUN: una herramienta didáctica para la derivación de programas funcionales**	Araceli Acosta (UNRC), Renato Cherini (UNC), Alejandro Gadea (UNC), Emmanuel Gunther (UNC), Leticia Losano (UNC), Miguel Pagano (UNC)
5883	**Metodología innovadora para el Estudio y Programación de Microprocesadores en Arquitectura de Computadoras**	Jorge R. Osio (UNAJ), Daniel Alonso (UNAJ), Eduardo Kunysz (UNAJ), Martín Morales (UNAJ)
5889	**Usando NDT como soporte a la enseñanza de programación web**	Yanina Medina (UNNE), Gabriel Osmar Pedrozo Petrazzini (UNNE), Cristina Greiner (UNNE), Gladys N. Dapozo (UNNE)

Experiencia de utilización de Herramientas Colaborativas para la enseñanza y el aprendizaje de la Programación de Computadoras

Lovos, Edith[1], Alejandro Gonzalez[2], Rodolfo Bertone[2]

[1] Universidad Nacional de Río Negro, Sede Atlántica
{elovos}@unrn.edu.ar
[2] Instituto de Investigación en Informática III-LIDI. Facultad de Informática,
Universidad Nacional de La Plata
alejandro.gonzalez@presi.unlp.edu.ar, pbertone@lidi.info.unlp.edu.ar

Abstract. En este trabajo se presentan algunos resultados y conclusiones preliminares sobre una experiencia de trabajo colaborativo apoyado en recursos tecnológicos provistos y/o compatibles con el EVEA Moodle. Las experiencias se aplicaron a un curso introductorio de enseñanza y aprendizaje de programación en la Lic. en Sistemas de la Universidad Nacional de Río Negro (UNRN) - Sede Atlántica. Se exponen algunas referencias teóricas que sostiene la propuesta y se realiza una descripción del contexto de aplicación. Finalmente se presentan los resultados de la implementación y conclusiones obtenidas.

Keywords: trabajo colaborativo, enseñanza, programación,

1 Enseñar y Aprender en colectivo

Maldonado Pérez [1] expresa la importancia de reconocer el carácter social que implica el enseñar y aprender en estos tiempos, donde el esquema convencional que posiciona al docente en el rol de enseñante y al alumno en su rol de aprendiz en forma exclusiva, ya no tiene lugar. Para la autora, el aprendizaje es un proceso social, construido a través de la interacción no solo del docente con los alumnos, sino entre alumnos y teniendo en cuenta el contexto y el significado que cada uno le asigna a lo que aprende. Esta forma de aprendizaje, responde a los postulados del psicólogo Jean Piaget, quien sostenía que el aprendizaje consiste en la generación de estructuras cognoscitivas que se crean a través de la modificación de los reflejos iniciales del recién nacido y que se van enriqueciendo a través de la interacción del individuo con el medio. A través de estas estructuras, el individuo adquiere información, usando los procesos de asimilación y acomodamiento de la misma. De esta forma, el proceso de aprendizaje no se basa en la memorización de la información, sino en asimilar o incorporar información a esquemas que poseen una información previa. El enfoque de Piaget se ve complementado, desde la perspectiva teórica de Vygotsky [2], que hacía énfasis en la interacción social como factor clave para el aprendizaje y la transmisión de cultura [1]. Según Johnson et al; [3], Vygotsky sostenía el carácter social del conocimiento y su construcción a partir de los esfuerzos cooperativos por

aprender, entender y resolver problemas. Un concepto clave, definido por Vygostky [2], es el de la zona de desarrollo próximo, entendiéndola como aquella zona situada entre lo que un estudiante puede hacer solo y lo que puede lograr si trabaja guiado por un instructor o en colaboración con otros pares más avanzados. Así, la enseñanza y como consecuencia el aprendizaje, sólo tiene lugar en la zona en la que el sujeto puede desarrollar una actividad en colaboración con otro [2]. En este sentido, Johnson [3], sostiene que a menos que los alumnos trabajen de manera cooperativa, no crecerán intelectualmente; por lo tanto, debe reducirse al mínimo el tiempo que los alumnos pasan trabajando solos en las actividades académicas. Maldonado Pérez [1], basándose en la teoría de Vygostky afirma que los procesos que desarrolla un grupo en interacción serán internalizados por cada uno de sus miembros, formando de esta manera parte de su propio aparato cognoscitivo. Por otra parte, destaca el espacio fundamental que ocupan los lenguajes y los procesos de comunicación en esta interacción. En cuanto al docente, la misma autora [1] señala que es su responsabilidad alentar, promover y crear el espacio adecuado que permita la construcción del conocimiento. En este sentido, se organizará la enseñanza y el uso de estrategias y metodologías apropiadas, que permitan la creación de nuevos espacios de interacción humana y tecnológica.

2 Ambientes Colaborativos & Enseñanza y Aprendizaje de la Programación

En las carreras vinculadas a la Informática, la enseñanza de la programación es una base fundamental y uno de los primeros cursos que deben tomar los alumnos ingresantes [4]. La enseñanza y aprendizaje de programación es una actividad intelectual compleja y dificultosa, tanto para los alumnos como para quienes llevan adelante la enseñanza; más aún cuando su impacto es muy importante en la mayoría de las asignaturas sucesivas y en el campo profesional del futuro egresado [5,6].

En el ámbito educativo, las actividades de aprendizaje colaborativas buscan desarrollar en los alumnos un conjunto de habilidades que se relacionan en forma directa con el objetivo que persigue la educación moderna, la formación en competencias que le permiten al alumno integrarse en una esta nueva sociedad mediada por tecnologías digitales, donde el docente desde su lugar debe ser, dinamizador, orientador y asesor de todo el proceso de enseñanza y aprendizaje [7]. En este sentido, Estévez [8], sostiene que los ambientes colaborativos pueden ofrecer un importante soporte a los alumnos durante las actividades aprendizaje de la programación. Y agrega que la resolución de problemas a través de la colaboración promueve la reflexión, un mecanismo que estimula el proceso de aprendizaje. Para el desarrollo de una actividad grupal los alumnos necesitan comunicarse, discutir y emitir opiniones a otros miembros del grupo, alentando de esta forma una actitud de reflexión que conduce al aprendizaje.

En cuanto a las características de una herramienta que esté orientada tanto para el aprendizaje como para el desarrollo colaborativo del software, algunos autores [9] señalan que deben estar incluidas: las actividades comunes, el entorno compartido y el espacio/tiempo. Por actividades comunes se entiende a aquellas tareas comunes que los participantes del grupo llevan a cabo; el entorno compartido brinda la posibilidad

de tener informado a cada miembro del proyecto sobre el estado de éste, lo que cada miembro está trabajando, etc.; y el espacio/tiempo soporta que la interacción del grupo de trabajo se produzca en el mismo lugar y momento. En cuanto a la interacción es posible encontrar dos tipos: síncrona o asíncrona, que a su vez puede ser distribuida o centralizada.
A continuación se describen las herramientas digitales que se utilizan en la experiencia que relata el artículo.

2.1 Virtual Programming Lab (VPL)

VPL es un producto de software de código abierto creado por el Departamento de Informática y Sistemas, de la Universidad de Las Palmas de Gran Canaria; que permite la gestión de prácticas de programación sobre el entorno virtual de enseñanza-aprendizaje (EVEA) Moodle[18], incorporando el ambiente de desarrollo de software al aula virtual de las materias donde se utiliza. Su arquitectura está compuesta de un módulo Moodle, un applet editor de código fuente y un demonio Linux que permite la ejecución remota de programas de forma segura. VPL tiene como propósitos el ahorro de tiempo y mejorar la gestión general de este tipo de actividades, tanto en los cursos de programación que se dictan en forma online como usando B-Learning, además de permitir la realización de las prácticas utilizando solo un navegador. La intención de la herramienta es facilitar el seguimiento y la orientación personalizada y continua del proceso de aprendizaje del alumno, contribuyendo de esta forma a tratar las dificultades a las que se enfrenta éste en la realización de las actividades de programación [11].
A nivel profesional, las herramientas comerciales que se utilizan para el desarrollo del software, presentan una amplia cantidad de opciones y de información que los alumnos que recién se inician en la práctica de la programación, no pueden comprender tan fácilmente porque aún no tienen los conceptos necesarios para manipularlas [10]. Así, VPL busca proveer a los alumnos novatos de un entorno de desarrollo que sea simple. Sus características más destacadas son: la posibilidad de editar el código fuente y ejecutar las prácticas de forma interactiva desde el navegador, ejecutar pruebas que revisen las prácticas y analizar la similitud entre prácticas para el control del plagio para algunos lenguajes de programación soportados [11]. La versión 2.0 de VPL incorpora características que permiten el trabajo en grupo. Así, cada grupo dispone de un repositorio compartido de entregas, donde cualquier integrante puede agregar una nueva versión del programa que están realizando y el resto del grupo recibirá el resultado de la evaluación.

2.2 Herramientas de Moodle: Foros y Wiki

Moodle dispone de una serie de herramientas que permite la colaboración dentro del aula virtual entre ellas los foros y wiki. A través de los foros, Moodle da lugar al planteo de debates y discusiones, posibilitando además la comunicación asincrónica. Los foros pueden estructurarse de diferentes maneras, y cada mensaje puede ser

evaluado por los participantes. Existen diferentes formas de visualizar los mensajes y los mismos permiten la inclusión de imágenes y adjuntar archivos. Cuando los participantes de un curso, se suscriben a un foro, recibirán copias de cada mensaje en su bandeja de correo. El participante con rol de profesor puede forzar la suscripción a todos los participantes.
En Moodle hay dos categorías de foros: Foro general (Se encuentra en la sección 0 del curso) y Foro de aprendizaje (Son foros de alguna sección específica del curso).
Una wiki es un espacio web colaborativo que puede ser editado por varios participantes, es decir todos pueden crear, modificar o eliminar contenido de forma interactiva; permitiendo así la escritura colaborativa [13] en [12]
Moodle dispone de la herramienta wiki, la cual se puede configurar al momento de crearla de un determinado tipo. Este tipo determina el ámbito de la misma y quien puede escribir y editar los cambios. Los tres tipos de wiki son: estudiante, grupo y profesor. La wiki puede funcionar en modo: sin grupos, grupos separados o grupos visibles al igual que los foros [14].

3 Actividades Prácticas Colaborativas

Se describe la implementación de la propuesta de enseñanza y de aprendizaje destinada a los alumnos ingresantes a la Licenciatura en Sistemas de la UNRN que tomen el curso de Programación I.

Programación I, es una materia perteneciente al área Algoritmos y Lenguajes de Programación; que se dicta en forma presencial en el primer cuatrimestre del primer año con un total de 96 horas. Tiene como objetivos generales que los alumnos puedan analizar problemas resolubles con computadora, poniendo énfasis en la modelización, la abstracción de funciones y en la descomposición funcional de los mismos, a partir de un paradigma procedural/ imperativo. Se realiza una introducción de las nociones de estructuras de datos, tipos de datos y abstracción de datos.

En cuanto a los alumnos, en su mayoría son ingresantes a la universidad, egresados recientemente del nivel medio, cuyas edades oscilan entre los 17 y 21 años, y que toman contacto por primera vez con la actividad de programación. Son varios los alumnos que llegan al curso con netbooks. Esto hace suponer que tienen cierto manejo de recursos tecnológicos como navegadores de internet y redes sociales tipo facebook entre otros. Por otra parte, es común verlos con sus teléfonos celulares navegando, escuchando música o mirando videos, aún dentro del espacio presencial de las clases.
El curso está dividido en clases teóricas y prácticas. En las primeras se desarrollan los conceptos teóricos previstos en el plan de estudio (resolución de problemas, estructuras de control, modularización, estructuras de datos) haciendo uso de ejemplos prácticos que permitan la aplicación de los conceptos analizados. Respecto a las clases prácticas, las mismas tienen como objetivo la aplicación de los conceptos trabajados en las clases de teoría, en la resolución de problemas computacionales, a través del diseño algoritmos. En un paso siguiente estas soluciones serán implementadas en un lenguaje de programación de alto nivel tipo Pascal. El énfasis

de la asignatura está puesto en la parte práctica, ya que para desarrollar la habilidad de resolver problemas usando algoritmos es fundamental el entrenamiento. Con este objetivo se diseñan actividades prácticas que enfrentan a los alumnos con situaciones problemáticas en las que tienen que decidir sobre la naturaleza del problema, seleccionar una representación que ayude a resolverlo (modelo) y, monitorear sus propios pensamientos (metacognición) y estrategias de solución [15].

El programa consta de seis unidades didácticas, cada una con su correspondiente trabajo práctico y tres Actividades Prácticas Entregables (APE) integradoras, las cuales deben ser entregadas y evaluadas para poder acceder al examen parcial. Las fechas de publicación de la APE, están establecidas en el cronograma de actividades de la materia.

Las APE consisten en la resolución colaborativa en equipos de trabajo, de problemas de mediana complejidad, cuya solución es un programa computacional que se implementará en el lenguaje de programación elegido por la cátedra. En el caso de Programación I, se utiliza el lenguaje Pascal.

La consigna de trabajo que se proponen a través de la APE, incluye la definición del problema, formas de entrega, consistente en un cronograma de actividades, fechas de previstas para cada etapa del proceso de resolución y recursos que se proponen para su desarrollo, información sobre pruebas, es decir se definen o se proporcionan los datos con los que serán puestos a prueba los programas y consideraciones especiales.

El desarrollo de las APE se propone que se realice combinando la forma de trabajo colaborativo y herramientas TIC, promoviendo de esta forma, la participación de los alumnos y el desarrollo de competencias transversales tales como el razonamiento crítico, la capacidad de análisis, el trabajo en equipo, la autorregulación y la comunicación. Teniendo en cuenta que los alumnos de Programación I son jóvenes que tienen cierto manejo de la tecnología, la intención de las APE es también que ellos las apropien como un recurso útil para construir y enriquecer su aprendizaje. Haciendo uso de las funcionalidades provistas por el entorno Moodle (Foro, Wiki, mensajería) y del laboratorio virtual de programación (VPL); se posibilita el desarrollo colaborativo del análisis y diseño de la solución y de la implementación del programa computacional que resuelve el problema propuesto en la APE. Las tres herramientas TIC se configuran de manera tal que cada grupo disponga de una instancia de las misma, de manera que los participantes solo puedan ver y editar las asociadas a su equipo.

Las consignas de las APE se presentan a través del aula virtual (archivo en formato .pdf), a la vez que se habilitan los espacios de Foro y Wiki y se indica el tutor asignado al grupo. El tutor es responsable de hacer el seguimiento del grupo y puede ser un docente de la teoría o de la práctica de la materia.

A continuación se describen las cinco etapas involucradas en el desarrollo de las APE; a saber:

Debate inicial: En la clase presencial siguiente a la presentación de la consigna en el aula virtual, se reserva una hora de la misma para que los grupos junto a los tutores asignados puedan debatir acerca del problema. Así, se busca atender dudas y

consultas sobre la consigna. Entre la fecha que se habilita la consigna de la APE sobre la plataforma y la fecha prevista para el debate, los estudiantes disponen de al menos 3 días para realizar una lectura crítica del problema en forma personal y grupal de manera de llevar a la clase de debate inicial dudas y consultas sobre la consigna propuesta.

Análisis y Diseño: en esta etapa se modela la solución al problema, el diseño modular y las estructuras de datos. Se propone utilizar una wiki y un foro ambas herramientas provistas por el entorno Moodle.

Implementación: en esta fase se propone continuar usando la wiki y el foro. Y se suma el laboratorio virtual VPL, con la opción de trabajo en grupo. A través de VPL, es posible que los grupos editen, compilen y ejecuten sus programas. Cada grupo tiene un repositorio compartido de entregas, y cualquier integrante del grupo puede entregar una nueva versión y todos los miembros de un grupo recibirán la misma devolución.

Presentación y defensa: en esta fase se propone la elaboración de una presentación que resume la solución al problema. La misma pueden subirla a la wiki de cada grupo y su exposición se desarrollará en una clase presencial de manera de poder compartir y debatir con el resto de los grupos las producciones realizadas; propiciando así la capacidad de comunicación.

Evaluación: La evaluación de las APE se desarrolla en tres partes: una evaluación del programa computacional a través del entorno virtual usando VPL, otra evaluación en forma presencial a modo de exposición y defensa de la solución propuesta y una tercera evaluación en forma de encuesta que permite que los alumnos evalúen el proceso de desarrollo de la APE, evaluando su propio desempeño, el del grupo y el del tutor asignado. De esta forma se propone evaluar la experiencia tomando en cuenta no sólo el resultado final de las APE - el programa computacional que resuelve el problema-, sino también el proceso de aprendizaje a nivel grupal e individual que dan lugar al mismo. Este proceso esta soportado a través del aula virtual de la materia y de las herramientas wiki, foros y VPL entre otras. Las evaluaciones de las APE servirán de información para los docentes y de orientación para el alumno. La evaluación que hacen los alumnos de sus compañeros de grupos, se apoya en la idea de un grupo de investigadores del Departamento de Informática y Sistemas Universidad de Las Palmas de Gran Canaria, España que entienden este tipo de evaluación como un complemento valioso que permite integrar al alumno en el proceso de evaluación del aprendizaje. De esta esta forma los alumnos pueden evaluar las competencias desarrolladas por sus pares durante el desarrollo de la actividad educativa. Los investigadores señalan que este tipo de evaluación requiere del alumno una mayor responsabilidad y el desarrollo de habilidades que le permitan valorar el trabajo de sus compañeros de equipo [16].

Respecto a la organización de los grupos de trabajo, en función de la complejidad que presentan las APE y teniendo en cuenta que desde los inicios de la carrera, en el año 2009, los alumnos inscriptos en el curso no supera los 50 en promedio, se propone que los equipos de trabajo no superen los 4 alumnos. En cuanto a su conformación, en esta experiencia se propuso para la APE1 que los alumnos decidan como

agruparse, luego en las siguientes APE los equipos fueron re-armados por el equipo docente de acuerdo al seguimiento realizado.

4 Resultados

El programa de la materia contempla el desarrollo de tres APE durante la cursada. Esta experiencia se inició con un grupo de 40 alumnos. De los cuales, para la APE1 se dividieron en 13 grupos de los cual completaron todas las etapas de la actividad 12 grupos. Para la APE2 se formaron 14 grupos y completaron la actividad 6 grupos. Para la APE3 se formaron 11 grupos y completaron la actividad 4 grupos. Respecto al desgranamiento que se observa, está asociado en parte con el hecho que muchos alumnos a medida que cursan las materias del primer año de la carrera, están también cursando las asignaturas introductorias de "Razonamiento y resolución de problemas" (RRP) e "Introducción a la lectura y escritura académica" (ILEA) que la UNRN ha dispuesto como un recorrido previo de ingreso universitario. De esta forma quienes no aprobaron estas materias antes del inicio del primer cuatrimestre, pueden cursarlas durante el mismo y/o rendirlas en forma libre en las fechas establecidas por el calendario académico.

Cómo se indicó anteriormente, al finalizar la fecha de entrega de cada APE los alumnos respondieron a una encuesta anónima, que permitió evaluar su propio desempeño, el de su grupo y el del tutor asignado. A continuación se exponen algunos resultados de las mismas. En cuanto a la autoevaluación los gráficos 1, 2 y 3, muestran el análisis de datos realizados hasta el momento. El gráfico 1 permite observar que los alumnos indicaron que su nivel de participación aumentó en cada APE. Solo para la primer APE, un poco más del 10% manifestó una participación nula. Cuando se les consultó a los alumnos acerca de porque no habían participado, algunos indicaron que por falta de coordinación y de organización. Otros manifestaron que no podían encontrarle utilidad al uso de la wiki o el foro, que preferían reunirse en forma personal.

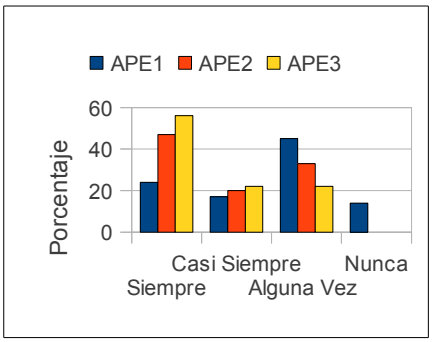


Gráfico 1 – Autoevaluación: nivel de participación en las e-actividades

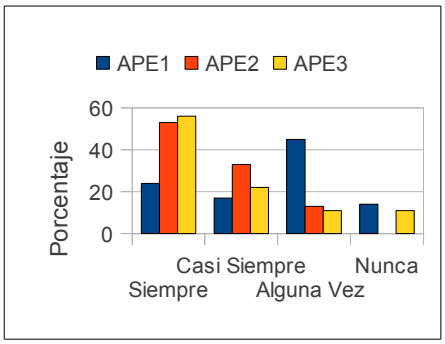


Gráfico 2 – Autoevaluación: dominio de los temas tratados

El gráfico 2, presenta el resultado de la autoevaluación que hicieron los estudiantes respecto a si tenían dominio (conocimiento y manejo) de la información que se discutía en cada APE. Se puede observar que en cada APE se produjo un incremento del mismo. En este sentido vale destacar que cada APE era integradora de los conceptos vistos en las unidades involucradas más los analizados en la APE anterior. Así por ejemplo para la APE2 el 53% manifiesta haber tenido siempre dominio de los temas tratados.
El gráfico 3, muestra las percepciones de los alumnos respecto a si el desarrollo de las APE les permitió una mejor comprensión de los conceptos involucrados. Se observa que más del 50% consideran que las APE contribuyeron en forma normal en las dos primeras actividades y en la última se observa que la contribución superó el 70%. Esta APE resultó una experiencia de investigación para los grupos, ya que la resolución del problema planteado requirió de conceptos no analizados en clase (Pilas y Colas).
Luego de finalizada la entrega de la APE 3 y tomado el examen parcial de la materia, se consultó a los alumnos acerca de si esta propuesta de trabajo les había permitido prepararse mejor para rendir el examen, aquí casi el 80% de los alumnos respondió positivamente, aún entre quienes indicaron no haber aprobado el examen.

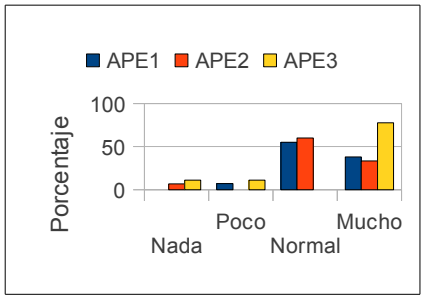


Gráfico 3 – Autoevaluación: comprensión de los conceptos

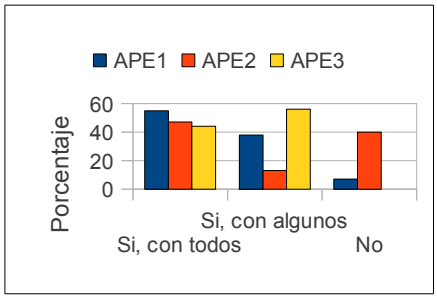


Gráfico 4 - Evaluación del grupo de trabajo

En cuanto a las evaluaciones que los alumnos realizaron sobre el grupo, en el gráfico 4 se observa que consultados acerca de si volverían trabajar con ese equipo, para las tres APE más del 40% respondió afirmativamente. Sólo en aquellos casos donde se observó a través del aula virtual y/o de la encuesta, que sus miembros no deseaban seguir trabajando juntos y/o que los alumnos lo manifestaron en forma personal equipo al docente, se hicieron cambios para la siguiente APE.
Ante la pregunta: ¿qué fue lo mejor y lo peor de trabajar en este grupo?, en general las respuestas coinciden y señalan la falta de dominio sobre los temas tratados. Esta situación ponía a los alumnos en distintos niveles, la despreocupación y la negativa de algunos integrantes a utilizar las herramientas TIC propuestas. En cuanto a lo mejor resaltaron el debate, la puesta en común de las posibles soluciones y el respeto hacia las opiniones de los demás. En la etapa de presentación y defensa de las APE, resulta importante destacar cómo evolucionaron las presentaciones digitales que realizaron los grupos y como los mismos interactuaron no solo con los docentes sino con los demás grupos, en la defensa oral de las actividades.

5 Conclusiones

Se presentó una propuesta de enseñanza de programación de computadoras basada en actividades prácticas colaborativas usando herramientas compatibles con Moodle. Para la primera implementación se puede observar que:

- Resultó difícil que los alumnos adoptarán el uso de foros y wiki, como un recurso para el desarrollo de las actividades didácticas específicas del área de programación, no así el uso del laboratorio virtual VPL. Para la mayoría de los alumnos ingresantes esta forma de trabajo y algunas de las herramientas TIC que la soportan, representan una experiencia novedosa. Así acordamos con otras investigaciones [17] que resaltan la necesidad de un tiempo de maduración de la misma por parte de los alumnos. Donde el mismo, debe estar en todo momento acompañado por la participación activa del tutor. En este sentido se propone a futuro el desarrollo de un curso pre-ingreso que trabaje sobre el uso de las mismas aplicadas al área específica de estudio.
- Hubo una evolución en el desarrollo de las presentaciones digitales y la calidad de las exposiciones a medida que avanzaban en la cursada.
- Desde los inicios de la Lic. en Sistemas en la UNRN Sede Atlántica en el año 2009, se puede observar que en las materias de programación de primer año el desgranamiento es muy alto de esta forma los alumnos que logran llegar al final del curso son muy pocos en relación a la cantidad de inscriptos. En este sentido la propuesta presentada, puede re-pensarse de manera que pueda convertirse en una herramienta que permita sostener a los alumnos a lo largo de la cursada. El desarrollo de actividades colaborativas en programación favorece el futuro desarrollo profesional de los estudiantes, se les presenta situaciones similares a las que van a tener que desarrollar.

Para la próxima implementación se trabajará en el ajuste de la propuesta teniendo en cuenta las observaciones de los estudiantes y docentes de la cátedra.

Como trabajo futuro se analizarán los tipos de problemas a resolver y su adecuación para el desarrollo de actividades colaborativas de programación.

Referencias

1. Maldonado Pérez, Marisel. El trabajo colaborativo en el aula universitaria. Revista Laurus, vol.13 nro. 23. Universidad Pedagógica Experimental Libertador. Caracas Venezuela. ISSN 1315-883X. (2007)
2. Vygotski, Lev. S. El Desarrollo de los procesos psicológicos superior. Barcelona . Grupo Editorial Grijalbo . (1978)
3. Johnson, David W., Johnson, Frank P. Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning. Needham Heights, MA: Allyn & Bacon. (1999)
4. Matthíasdóttir, Á. How to teach programming languages to novice students? Lecturing or not?, Proceedings of the International Conference on Computer Systems and Technologies, June 15-16, University of Veliko Tarnovo, Bulgaria (2006).
5. Costelloe, E. Teaching Programming. The State of the Art. Department of Computing, Institute of Technology Tallaght, Dublin 24. CRITE Technical Report, 2004a. (2001)

https://www.scss.tcd.ie/disciplines/information_systems/crite/crite_web/publications/sources/programmingv1.pdf Abril 2012
6. Lahtinen E, Ala-Mutka K, et al. A Study of the Difficulties of Novice Programmers. 10Th annual SIGCSE conference on Innovation an technology in computer science education ItiCSE '05Linder, et al, (2001) "Facilitating Active Learning With Inexpensive Mobile Robots". Journal of Computing in Small Colleges,16, 4. (2005) http://delivery.acm.org/10.1145/380000/378656/p21-linder.pdf?key1=378656&key2=7062133701&coll=guide&dl=acm&cfid=15363060&cftoken=16364368 . Junio 2012
7. Filippi, J. L, Lafuente, J., Bertone, R. .Diseño de un Ambiente de Aprendizaje Colaborativo. En actas del V Congreso de Tecnología en Educación y Educación en Tecnología. Universidad Nacional de la Patagonia Austral. Calafate. (2010) .
8. Esteves M., Morgado L., Martins P., Fonseca B. "The use of Collaborative Virtual Environments to provide student's contextualisation in programming". En: Proceedigns of m-ICTE 2006. (2006)
9. González de Rivera Fuentes, M., Paredes Velasco, M. Aprendizaje con programación Colaborativa. Número 2008-02. Serie de Informes Técnicos DLSI1-URJC. ISSN 1988-8074. Departamento de Lenguajes y Sistemas Informáticos I Universidad Rey Juan Carlos (2008).
10. Pérez Pérez, J. R., Paule Ruiz, J.M., Del Puerto M., Cueva Lovelle J. M. Capítulo 3. Sistemas orientados a la mejora de la calidad del software. En congreso IV International Conference on Multimedia and Information & Communication Technologies in Education (m-ICTE2006). (2006).
11. Rodriıguez del Pino, J.C., Royo Rubio E., Hernandez Figueroa. VPL: Laboratorio virtual de programación para Moodle. En Actas de las XVI Jornadas de Enseñanza Universitaria de Informática, Jenui 2010, pags. 429–435, Santiago de Compostela, Julio 2010. http://www.di.uniovi.es/~juanrp/investigacion/tesis/2%20Tesis_SICODE_Estado_del_arte.pdf Junio 2012
12. Salinas Silvia. Software para trabajo colaborativo y bibliotecas. E-LIS. E-prints in Library and Information Science; 8 2008 http://eprints.rclis.org/14721/1/Software_para_Trabajo_Colaborativo_y_Bibliotecas1.pdf
13. Neri, C. Fernández Salazar, D. Telarañas de Conocimiento. Educando en tiempos de la web 2.0. ISBN 978-987-1426-01-0. (2008).
14. Moodle 1.9. WIKI. Unitat de Suport Tecnicopedagògic - CAMPUS EXTENS Universitat de les Illes Balears. Edifici Aulari. (Illes Balears) (2010). http://campusextens.uib.es/_doc/bonespract/proyec-manuales/castellano/wiki.pdf
15. García, Juan Carlos. Solución de Problemas mediante la programación. Portal Eduteka . http://www.eduteka.org/modulos.php?catx=9&idSubX=298 Junio 2013
16. Diaz Roca, M. Rodriguez del Pino, j.C., Hernández Figuero, Z., Vicente, C. M. El Gestor de Coevaluacion Orientado Grupos. Una herramienta de apoyo a la participación del alumno en el proceso de evaluación. 7ª Conferencia Ibérica de Sistemas y Tecnologías de Información. CISTI 2012, Madrid. (2012) http://www2.dis.ulpgc.es/~mdiaz/GestorCoevaluacionOrientadoGrupos.
17. Lillo Zuñiga, Felix Revista de Psicología Universidad Viña del Mar (2013) Vol. 2, Nº 4 109 – 142. ISSN http://sitios.uvm.cl/revistapsicologia/revista-detalle.php/4/25/contenido/aprendizaje-colaborativo-en-la-formacion-universitaria-de-pregrado
18. Moodle. https://moodle.org/

Enseñanza de técnicas de elicitación de requerimientos

Alejandro Oliveros, Javier Zuñiga, Sergio Corbo, Patricia Forradellas, Sandra Martínez

{aoliveros,sjzuniga,samartinez}@uade.edu.ar,
sergio.corbo@gmail.com,psforrade@hotmail.com
INTEC – UADE, Lima 775, CABA, Argentina

Abstract. La enseñanza de las técnicas de elicitación de requerimientos posee especiales dificultades por la imposibilidad de generar un contexto equiparable al mundo real. La *observación* constituye una poderosa herramienta de aprendizaje que se propone utilizar para enseñar a estudiantes de un primer curso de Ingeniería de Requerimientos las técnicas de entrevistas. En esta comunicación se informa un experimento desarrollado para evaluar la calidad y organización de las preguntas en una entrevista con usuario para obtención de requerimientos.

Keywords: Elicitación de requerimientos, entrevistas, observación, enseñanza

1 Introducción

El presente artículo continua con la experiencia reportada en [1] y [2] en los que se encuentran desarrollados los fundamentos del presente experimentos

1.1 Las entrevistas como técnicas de elicitación

En [2] se encuentran los detalles y fundamentos acerca de las entrevistas. El resultado del proceso de elicitación de requerimientos es el conocimiento necesario para producir el modelo de requerimientos de un dominio de problema dado [3]. La más sencilla forma de interacción es la "open-ended interview" [4]. Estás técnicas requieren habilidades especiales del analista [3], [4]. Este tipo de entrevistas proviene de prácticas previas de la Ingeniería de Software y Sistemas de Información.[5] Las entrevistas no estructuradas son ventajosas en cuanto a efectividad y completitud del output [6].

Con el objetivo de mejorar la enseñanza de las técnicas de entrevistas se desarrollaron varias experiencias en un curso de grado. La pregunta que guió esta

experiencia fue: ¿resulta de utilidad la observación como técnica para la enseñanza de entrevistas?

1.2 El problema de enseñar técnicas de elicitación de requerimientos

Existen varios problemas para la enseñanza en las aulas de las técnicas de entrevista[2]:

1. dificultad de ejecutar una práctica real:
2. subestimación por parte de personas de formación tecnológica de las técnicas "blandas" que requiere la elicitación de requerimientos.

La multiplicidad de abordajes que existen para enfrentar este problema pone de manifiesto que está lejos de su solución [2]. Nuestro abordaje intenta utilizar la observación como técnica de aprendizaje través del uso de una "cámara Gesell" (*Gesell dome* en inglés), más detalles en [2]

Este trabajo está organizado de la siguiente forma. En el punto 2 se resumen algunos puntos clave del enfoque de aprendizaje propuesto, en el punto 3 se reproduce el contexto de la experiencia. El punto 4 describe la experiencia con detalle. En el punto 5 se evalúa comparativamente los resultados obtenidos en los dos casos de estudio. Por último se plantean algunas conclusiones y trabajos futuros.

2 La observación como método de aprendizaje

Sumariamente, el enfoque propuesto se basa en que el *aprendizaje* se propone conseguir un cambio permanente en la conducta del individuo atribuible a una experiencia [7]. El aprendizaje concluye en un cambio en la conducta [8]

"En términos generales, por aprendizaje cognoscitivo se entiende el conocimiento, el saber, el anticipar o utilizar en otra forma los procesos mentales superiores ricos en información. El aprendizaje cognoscitivo va más allá del condicionamiento básico, pues abarca la memoria, el pensamiento, la resolución de problemas y el lenguaje." [7] Las investigaciones de Albert Bandura en el campo de las teorías de la personalidad contribuyeron a la constitución del campo del "Social Learning" como un desarrollo de las teorías cognitivas del aprendizaje [9] y han conformado una de las principales corrientes de las teorías del aprendizaje [10], [11]. El *aprendizaje por observación* se produce al exhibir comportamientos derivados de la exposición a conductas modeladas. Este aprendizaje consta de cuatro pasos [11], [12]: *atención, retención, reproducción* y *motivación*.

3 El contexto de la experiencia desarrollada

La materia Ingeniería de Requerimientos integra los planes de estudio de dos carreras, Ingeniería en Informática y Licenciatura en Informática, ambas de cinco años de duración. El curso se dicta en el primer cuatrimestre del segundo año. En el primer año de estudios hay un curso de Análisis Estructurado, aunque no es una exigencia

haberlo aprobado para cursar Ingeniería de Requerimientos. El libro de texto es el de Wiegers [13] y además se utiliza material de Procees Impact [14]. Más detalles sobre la materia se encuentran en [2]

4 Descripción de la experiencia

4.1 Ideas básicas del proyecto

La investigación se desarrolló siguiendo los estándares de la investigación experimental, ´para ello se elaboró un documento con el detalle de las actividades a realizar en el proyecto y una descripción de los productos a obtener y el registro de los pasos datos.

El esquema básico del proyecto es que equipos de 3 a 4 alumnos observan a otro equipo de similar tamaño que realiza una entrevista a un usuario en una Sala Gesell.

La entrevista al usuario se hizo en la Sala Gesell para que puedan observar su desarrollo los restantes equipos. Observaron dos entrevistas con usuarios pertenecientes a diferentes dominios de aplicación

La entrevista forma parte del trabajo final de la materia, vale decir: los alumnos utilizan los requerimientos obtenidos como insumo para elaborar la especificación de un sistema. Con este enfoque se trató de conseguir mayor motivación de los alumnos por la actividad.

La experiencia se hizo en el curso del turno mañana del primer cuatrimestre de 2013. Se conformaron equipos de 4 a 5 alumnos y un grupo hizo la entrevista personal dentro de la Sala Gesell.

Las entrevistas dentro de la sala se filmaron para que los investigadores (toda la cátedra de la asignatura) puedan evaluarlas y así cotejar sus evaluaciones con las de los alumnos.

A fin de homogeneizar el desempeño de los alumnos, todas las consignas se trasmitieron sobre la base de documentos escritos especialmente para este proyecto indicándose a los alumnos que debían ser seguidos por todos los equipos. Los alumnos ejecutaron las entrevistas y evaluaciones sobre la base de ese material elaborado especialmente sobre el tema entrevistas.

Previstamente a ejecutar cada una de las entrevistas los alumnos establecieron el alcance de los requerimientos a identificar (alcance del sistema)

El contexto general del trabajo de los equipos entrevistadores fue que en la entrevista debería obtenerse conocimiento a volcar en una lista de requerimientos elaborado según las pautas de la cátedra. Para ello los estudiantes recibieron indicaciones acerca de registrar objetivos, necesidades, expectativas y requerimientos que surjan de las entrevistas. Con este objetivo los estudiantes deberían formular las preguntas adecuadas. En el presente artículo se informa la evolución de la calidad de las preguntas y el comportamiento de los alumnos durante la entrevista en sí misma.

4.2 Descripción del primer caso: sistema de becas

La experiencia se desarrolló en los meses de marzo y junio de 2013 (desde el 27 de marzo). En lo que sigue se describe siguiendo la secuencia de las clases en las que se desarrollaron las actividades

Clase de 27 de Marzo. Los docentes del curso presentan a los alumnos la iniciativa de realizar un trabajo de investigación acerca de la enseñanza de requerimientos asociad con el desarrollo del curso. Se explica la dinámica de organización del proyecto y el papel de los grupos. La iniciativa tiene buena aceptación por parte de los alumnos, se organizan 9 equipos de trabajo cada uno de ellos compuesto por 4 a 5 alumnos.

Clase 3 de Abril. Se desarrollan actividades indicando la importancia de comprender el dominio en el proceso de elicitación. Hacen un ejercicio en clase de describir un dominio. Se presenta el LEL (Léxico Extendido del Lenguaje) como una herramienta de comprensión del vocabulario del dominio, se trasmiten algunas sugerencias sobre cómo aplicarlos en las actividades prácticas relacionadas con proyecto. El LEL se presenta como una herramienta para ayudar a entender el dominio en el contexto del proceso de elicitación. Se introduce el marco a la primera actividad práctica del proyecto en la clase siguiente (asistencia a la clase de un usuario del Dpto. de Becas de UADE, quien dará las primeras visiones acerca del alcance de una problemática de negocios).

Clase 10 de Abril. Se desarrolla una clase teórica sobre Técnicas de Entrevistas de acuerdo con el contenido del material disponible por los alumnos, se hace foco en aquellas técnicas que podrán ser de utilidad en la primera entrevista a un usuario del Dpto. de Becas de UADE. Se hace hincapié en la importancia de poder elaborar y fijar una primera versión del alcance de la necesidad que planteara el usuario, como entregable de la primera entrevista. El sistema en consideración es el Sistema de Legajos del Depto. de Becas de la Universidad. Se realiza la 1er entrevista con un usuario del Dpto. de Becas de UADE, quien asiste invitada a la clase, en la misma todos los alumnos formulan preguntas cuyos principales objetivos eran:

1. comprender el dominio en el que se desarrolla la necesidad
2. lograr una primera versión del alcance de la necesidad planteada

La duración de la entrevista en el curso se desarrolla en aproximadamente 1 hora y comienza con la presentación del entrevistado a los fines de la experiencia corresponde mencionar que se trata de la Jefe del Depto. Becas, que posee formación en el área de informática y que posee una amplia experiencia en el dominio. Al finalizar la clase se les trasmite la consigna de elaborar por equipos las preguntas que deberían hacerse al usuario en la siguiente entrevista (por un grupo y en la Sala Gesell).

La observación directa por parte de los docentes presentes en el curso (tres) concluyó en varias observaciones.

- Si bien el foco era el alcance y límites del sistema, destinaron muchas preguntas a detalles. Ejemplo: “¿existen cupos para las becas?, ¿cuántos?, ¿de qué tipo? (que parece reflejar más un interés personal que profesional).

- Los docentes debieron intervenir en dos oportunidades para orientar hacia el foco. Ejemplo: "¿Por qué motivos se puede consultar un legajo archivado?".
- Preguntas con múltiples interrogantes. Ejemplo: ¿Cómo está conformado el legajo de becas?, ¿Se lleva un historial del mismo?, ¿Durante cuánto tiempo se archiva?"
- En algunos casos se observó un comportamiento agresivo hacia el entrevistado. Ejemplo: interrogando sobre afirmaciones en apariencia contradictorias entre si

Clase del 17 de Abril. Los alumnos entregan a los docentes las preguntas elaboradas por los equipos, se separan todas las preguntas y se agrupan por tema. En forma conjunta se seleccionan las preguntas que se utilizaran en la Sala Gesell para la entrevista. Los docentes indicaron que las preguntas sobre el proceso global las hicieran el principio, el resto lo ordenaron los entrevistadores. Sobre la base de la calidad de las preguntas presentadas, los docentes seleccionan a los entrevistadores. Luego se dirigen a la Sala Gesell y el grupo de alumnos seleccionados realiza las entrevistas siguiendo la selección de preguntas hecha previamente, la entrevista es filmada. Los restantes alumnos observan fuera de la sala.

La observación de los docentes (tres) de la entrevista concluyó:

- Las preguntas son formuladas correctamente.
- Varias preguntas fueron muy genéricas, evitaron hacer preguntas más puntuales y específicas (puede ser una reacción a acotaciones hechas en clase).
- Cuando el usuario repreguntaba, formulaban las preguntas en los mismos términos que la primera vez, no permitiendo aclara mucho.

4.3 Descripción del segundo caso: UADE Arts

Esta caso está descripto con detalle en [2].

Clase del 19 de junio. Los estudiantes se reúnen en la sede del UADE Art. La reunión tiene el mismo formato que la del 10 de Abril con la responsable de becas. El usuario es muy ordenado y preciso en sus respuestas. La conclusión de la observación directa por parte de los docentes es que las preguntas carecen de las deficiencias que se observaron en el caso anterior.

Clase del 26 de junio. Igual formato que la clase del 17 de abril. Para hacer de entrevistadores se seleccionan alumnos distintos de la primera vez. A los entrevistadores, se les sugiere que previamente a la entrevista hagan una reunión de coordinación de la dinámica de la reunión.

Sobre esta entrevista los docentes concluyeron:

- Realizan bien las preguntas.
- Cuando el usuario no comprende la pregunta la reformulan de otra manera.
- Se observa una buena coordinación entre los entrevistadores.
- El flujo de la entrevista es adecuado

5 Evaluaciones

Disponemos de varias evaluaciones de las entrevistas en la Sala Gesell. La **primera** por los tres docentes del curso de los que fuimos reflejando a lo largo de la descripción de la experiencia y que fue formulada inmediatamente de producida la entrevista. Esas conclusiones se pueden resumir en: erradicación de las preguntas genéricas, flujo adecuado, correcta reformulación de las preguntas.

La **segunda** fuente de evaluaciones es la entrevista sostenida por al primer autor del presente (que no es docente del curso) con dos de los tres docentes. De ella se concluyó que en la segunda experiencia:

- mejoró la forma de preguntar (no hay múltiples interrogantes)
- lenguaje más cercano al usuario
- la coordinación entre los entrevistadores hubo que inducirla por los docentes
- valorizaron la secuencia que fue sugerida por los docentes: "primero entender el proceso general y luego los detalles"
- mejora de las repreguntas

En comparación con otras experiencias [2] la discusión en conjunto y depuración de las preguntas ayudo mucho a la mejora de la calidad de las preguntas.

La **tercera** fuente de evaluaciones consistió en que dos docentes de la cátedra, que no pertenecían al curso del experimento, observaron los videos de las entrevistas en la Sala Gesell y realizaron una comparación entre ambas. El análisis realizado se resume en el Cuadro 1.

Cuadro 1. Comparativo de entrevistas

Entrevista Becas	Entrevista UADE Art
Al inicio de la entrevista se revisa vagamente el alcance del sistema.	Al inicio de la entrevista se revisa y confirma el alcance y límites del sistema
Entrevistadores con escaso conocimiento del dominio	Entrevistadores con un buen conocimiento del dominio
Preguntas no muy claras	Preguntas claras y concretas
Preguntas generales del proceso	Preguntas generales y de detalle
Secuencia de preguntas desordenada	Secuencia de lo general a lo particular
Preguntas con tecnicismos	En general sin tecnicismos
Sin referencias a la primera entrevista	Referencia a la primer entrevista
Preguntas con supuestos (incluso incorrectos)	Las preguntas no incluyen supuestos
No se valida que forma parte del sistema y que no	Preguntas que validan que se va a incluir en el sistema
En el final no se valida lo relevado	En todo momento se valida lo relevado
Adecuado uso del tiempo	Adecuado uso del tiempo
Los tres entrevistadores utilizan el "Usted".	No todos utilizan el "Usted"
Presionados por las preguntas escritas	Presionados por las preguntas escritas

El Cuadro 1 resulta auto explicativo, pero cabe mencionar con un elemento a revisar es el atarse exclusivamente al libreto de las preguntas escritas.

Al final de las entrevistas se consultó a los alumnos y esta **cuarta** fuente consideró a la experiencia como muy real, incluso por aquellos alumnos que ya trabajan en informática, además le resultó muy motivador utilizar un recurso como la Sala Gesell.

6 Conclusiones y trabajos futuros

Sobre la base de un esquema conceptual de aprendizaje presentado en [2], se continuó el desarrollo de la experiencia de aprendizaje de técnicas de entrevistas. Estas tienen varias componentes, uno de ellas la forman sus preguntas y la dinámica de su desarrollo. Sobre la base de una experiencia realizada en un curso de Ingeniería de Requerimientos de 2do año de la carrera de Ingeniería Informática se investigó la forma de enseñar a desarrollar entrevistas.

La efectividad de ese enfoque de aprendizaje estará dada por la capacidad los alumnos para ejecutar una entrevista adecuada. En este trabajo formulamos la idea de adecuación de la entrevista en términos de evaluaciones que se pueden hacer sobre su ejecución y comparando la calidad de las preguntas formuladas.

La conclusión general, más allá de las que fueron detallándose en el texto, es que visualizar las entrevistas realizadas por sus compañeros resulta motivador para los alumnos y un elemento disparador de mejoras. La evolución entre la primera y la segunda experiencia así lo demuestra.

Pero el criterio de calidad definitivo de una entrevista son sus resultados, en nuestro caso esos resultados son los requerimientos. El experimento realizado incluye disponer de los requerimientos obtenidos y justamente esa evaluación de los requerimientos obtenidos en términos de completitud de los términos tratados con relación al sistema en construcción.

7 Referencias

[1] A. Oliveros, J. Zuñiga, R. Wehbe, S. Rojo, y S. Martinez, «Enseñanza de elicitación de requerimientos», presented at the WICC 2012 - XIV Workshop de Investigadores en Ciencias de la Computación, Posadas - Misiones, 2012.

[2] A. Oliveros, J. Zuñiga, R. Wehbe, S. Rojo, y F. Sardi, «Enseñanza de elicitación de requerimientos», presented at the Congreso Argentino de Ciencias de la Computación (CACIC), Bahía Blanca, 2012.

[3] P. Loucopoulos y V. Karakostas, *Systems Requirements Engineering*. McGraw-Hill, 1995.

[4] J. A. Goguen y C. Linde, «Techniques for requirements elicitation», in *Requirements Engineering, 1993., Proceedings of IEEE International Symposium on*, San Diego, CA , USA, 1993, pp. 152 - 154.

[5] B. Nuseibeh y S. Easterbrook, «Requirements Engineering: A Roadmap», in *ICSE '00 Proceedings of the Conference on The Future of Software Engineering*, Limerick, Ireland, 2000, pp. 35 - 46.
[6] O. Dieste y N. Juristo, «Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques», *IEEE Transactions on Software Engineering*, vol. 37, n.º 2, pp. 283-304, abr. 2011.
[7] D. Con, *Psicología*. México: International Thomson Editores, 2005.
[8] F. Rojas Velásquez, «Enfoque sobre el aprendizaje humano», Departamento de Ciencia y Tecnología del Comportamiento. Universidad Simón Bolívar, jun. 2001.
[9] A. Bandura, *Social Learning theory*. Englewood Cliffs, N.J.: Prentice Hall, 1977.
[10] F. Ashworth, G. Brennan, K. Egan, R. Hamilton, y O. Sáenz, «Learning Theories and Higher Education», *Level3*, vol. 2, jun. 2004.
[11] D. H. Schunk, *Teorías del aprendizaje*, 2da edición. México: Prentice-Hall, 1997.
[12] C. G. Boeree, «Personality Theories», *Boeree Home Page*. [Online]. Available: http://webspace.ship.edu/cgboer/perscontents.html. [Accessed: 26-dic-2011].
[13] K. Wiegers, *Software Requirements*, 2nd ed. Microsoft Press, 2003.
[14] «Process Impact». [Online]. Available: http://www.processimpact.com/.

Generando Entornos de Investigación y Desarrollo utilizando Redes Inalámbricas de Sensores (WSN)

Eduardo O. Sosa[1], Diego A. Godoy[2], Edgardo A. Belloni[2]

[1] Facultad de Ciencias Exactas, Químicas y Naturales - Universidad Nacional de Misiones. Félix de Azara 1552. N3300LQH. Posadas, Argentina.
[2] Departamento de Ingeniería y Ciencias de la Producción - Universidad Gastón Dachary, Av. López y Planes 6519, N3301BOL. Posadas, Argentina

eososa@unam.edu.ar,{diegodoy,ebelloni}@ugd.edu.ar

Resumen. Las Redes Inalámbricas de Sensores (WSN) jugarán un papel preponderante tanto en las actividades académicas y como en el desarrollo en nuestro país. En este artículo se reportan actividades desarrolladas en dos universidades argentinas, utilizando a la tecnología WSN como disparador de actividades de investigación, desarrollo. La aplicación de ésta tecnología de última generación se enmarcan en el dominio de "Internet de las Cosas" y "Ciudades Inteligentes". Las actividades teórico-prácticas desarrolladas introdujeron a un número cada vez mayor de interesados en la experimentación, hacia el desarrollo de soluciones frente a situaciones de la vida cotidiana, buscando soluciones utilizando WSN. Se presentan resultados de proyectos de la vida real, por lo que amerita considerar la implementación del estudio de las WSN en las curriculas de las carreras con contenido de las Tecnologías de la Información y Comunicación.

Palabras clave: Sensores inalámbricos; Redes ad hoc; Internet del Futuro; Internet de las Cosas.

1 Introducción y Propósitos

Las redes ad-hoc son sistemas sumamente complejos. En ellos conviven y participan muchos conceptos, protocolos, tecnologías, algoritmos, y elementos que deben ineludiblemente trabajar conjuntamente. La aplicación de las redes móviles ad hoc (MANET) y las WSN es sumamente diversa, yendo desde pequeñas redes estáticas limitadas en su existencia por la necesidad de disponibilidad de energía, a redes a gran escala con gran dinámica y mucha movilidad. Si bien los nodos sensores han sido utilizados desde hace décadas, el desarrollo de la tecnología ha sido exponencial a partir de 1998, con el proyecto SmartDust.

Las redes de sensores inalámbricos, proveen una tecnología que permite operar de forma autónoma a cada uno de los nodos, sin depender de infraestructura alguna. Son parte de aquellos objetos cooperantes, que residiendo en el dominio de la computación ubicua; permite desarrollar una gran variedad de aplicaciones prácticas. La

computación ubicua es un modelo de interacción de personas y equipos, en los cuales el procesamiento ha sido asimilado invariablemente a los elementos y actividades del día a día. Cada uno de los objetos con los cuales interactuamos, tienden a ser integrados con sensores de alguna naturaleza. Son redes auto-configurables de pequeños nodos desplegados en cantidades suficientes de tal manera de interaccionar con el mundo. Es posible mencionar casos de éxito como el monitoreo de hábitats naturales de aves, control de las migración de animales, vulcanología, contralor de parámetros en viñedos, calidad de vida de los internos en residencias geriátricas, eventos deportivos multitudinarios, y diversos otros dominios y entornos. La solución implementada en cada caso ha sido más competitiva que las existentes, ya que se puede implementar como red fija o móvil.

Computación ubicua es un modelo posterior de la interacción persona-computadora de escritorio, y es un término ya utilizado por Mark Weiser en los años 1990's [1]. Hoy las WSN forman la columna vertebral de una nueva Internet, principalmente ubicua, como parte indivisible de "Internet de las cosas (IoT)" [2], dominio en el cual cada "cosa" existente en el mundo físico también puede convertirse en un elemento que está conectado a Internet. Las "cosas" se pueden caracterizar como pequeños dispositivos capaces de diferentes tareas "inteligentemente".

Si a la brecha digital se la define como la diferencia entre los que tienen acceso regular y eficaz de las tecnologías digitales y los que no, entonces la brecha científica se puede definir como la brecha entre quienes tienen acceso a los datos científicos y los que no. Estamos persuadidos que el uso de WSN en los países en vías de desarrollo puede ayudar a llenar este vacío en el estado del arte y de la técnica. Abogamos por la utilización de WSN para el desarrollo en el dominio de las ICT4D (ICT for Development) ya que se convierte, aplicándose a diversos escenarios [3], en una herramienta válida para reducir la brecha científica y tecnológica existente.

La implementación de las WSN aporta nuevas fortalezas al diseño curricular de carreras en Informática, agregando valor actualizado para los propósitos educativos en aquellas instituciones académicas que avancen sobre el tema.

En nuestro país existen hoy 2,9 investigadores, tecnólogos y becarios por cada 1.000 personas de la población económicamente activa. Se espera aumentar a 4,6 en 2020, según el escenario más pesimista [4], logrando duplicar la cantidad de científicos en 7 años.

Uno de los ejes principales del plan es la "focalización" en sectores que se consideran estratégicos, la agroindustria, el ambiente y el desarrollo sustentable, el desarrollo social, las energías, la industria, y la salud. En todos esos ambientes es esperable que las WSN jueguen un rol sumamente importante.

La capacidad de realizar mediciones directas y determinar las estrategias de reconocimiento y caracterizar patrones; conjuntamente con una acertada explotación de los recursos computacionales disponibles en la tecnología; representan retos ingenieriles muy interesantes de abordar en el ámbito científico académico. Para validar la tecnología se hace necesario un amplio portafolio de aplicaciones como una prueba del concepto. Para ello las redes desplegadas deben adecuarse al medio bajo estudio e investigación, debiendo considerarse no solo el impacto científico potencial, sino también el impacto en la sociedad. Las bondades y potenciales aplicaciones de las

WSN, coadyuvará en la adhesión de mayor cantidad de voluntades, siendo fundamentales aquellas que se encuentren relacionadas con actividades de desarrollo de hardware.

2 Génesis y Contexto

La Facultad de Ciencias Exactas, Químicas y Naturales (FCEQyN) de la Universidad de Misiones (UNaM), el Centro de Investigación en Tecnologías de la Información y Comunicaciones (CITIC) de la Universidad Gastón Dachary (UGD), y el Parque Tecnológico Misiones (PTMi) han trabajado activamente en la difusión y promoción de actividades centradas en tecnologías de bajo costo y alto impacto en la industri y la sociedad, siendo uno de los objetivos la formación de personas que puedan convertirse en propagadores de los conocimientos y las capacitaciones recibidas.

Por convenios con la Universidad de Lübeck (Alemania), el LINTI –Laboratorio de Investigación de Nuevas Tecnologías Informáticas de la Universidad Nacional de La Plata y la FCEQyN se ha desarrollado el Proyecto "Hacia una Red Global de Sensores Interconectados. Un ensayo experimental Argentino-Alemán", aprobado en el marco del Programa de Cooperación Científico-Tecnológico entre el Ministerio de Ciencia, Tecnología e Innovación Productiva de la República Argentina (MinCyT) y el Bundesministerium für Bildung und Forschung (BMBF) de Alemania. La meta principal del proyecto ha sido obtener resultados valederos tanto desde el punto de vista práctico como desde lo teórico y académico, promoviendo cursos y actividades a nivel de grado y postgrado, los que no son tan comunes en las actuales curriculas de las universidades argentinas. El carácter multidisciplinario de la actividad involucrada en el proyecto ha logrado aportes significativos al estudio y modelización de tráfico en las redes ad-hoc, como así también lo concerniente a encaminamiento en las citadas redes. En éste marco se realizaron capacitaciones sobre WSN en la UNLP y la UNaM impartidas por científicos y académicos alemanes.

De la actividad participaron 10 estudiantes de postgrado de la UNLP, otorgando 3 créditos válidos para Maestrías y Doctorados. En la UNaM tomaron dicho curso 12 personas de las carreras de Informática e Ingeniería Electrónica de la UNaM. Como corolario de la actividad, se ha defendido una tesis de Doctorado en Ciencias Informáticas en la UNLP [5] en el año 2011.

De la sinergia establecida entre grupos de investigación en la UNaM y la UGD, se han implementado charlas orientadas a docentes y alumnos de la carrera Ingeniería en Informática denominándoselas "Internet de las Cosas e Inteligencia Ambiental". En ese contexto se concretaron los talleres "Programación de WSN" y "Hacia la Internet del Futuro, Programando WSNs" y "Redes de Sensores Inalámbricos e Internet del Futuro". Estas actividades incluyeron importante carga horaria en actividades prácticas, que ha sido posible abordar por disponer un Laboratorio desplegado con una WSN de 15 nodos.

3 Talleres Desarrollados

Los objetivos, dinámica, contenidos, práctica, tecnologías utilizadas, y algunas consideraciones relativas a la evaluación de los talleres se indican a continuación.

3.1 Objetivos

Objetivo General.

- Concientizar sobre el potencial de la tecnología WSN, insistiendo en el hecho consisten en dispositivos de bajo costo; resultando apropiada fundamentalmente para aplicaciones en el medio ambiente.

Objetivos Particulares.

- Proporcionar comprensión general sobre ésta nueva tecnología y el nuevo paradigma de las redes data-céntricas.
- Apreciar de la naturaleza interdisciplinaria de WSN revelando sus potenciales aplicaciones para la región.
- Inculcar habilidades prácticas a través de la auto-motivación, la formación práctica no rutinaria, y de actividades de diseño en equipo, utilizando pensamiento crítico, trabajo en equipo y habilidades de comunicación.
- Desarrollar una estructura sostenible generando la infraestructura necesaria para formar a una futura generación de capacitadores, capaces de interactuar a nivel local, para así coadyuvar al desarrollo tecnológico.
- Fomentar con un enfoque regional, el desarrollo de un sentido de comunidad entre los participantes, despertando el interés por la aplicación de la tecnología WSN como una herramienta válida para resolver problemas locales.
- Promover la conformación de un grupo de trabajo sobre WSN en la UNaM y la UGD integrado por investigadores, profesionales y estudiantes.
- Elaborar documentos técnicos sobre dispositivos sensores
- Dictar charlas y conferencias de sociabilización en distintos estamentos de nuestra comunidad.

3.2 Dinámica y Contenidos Tratados

Los talleres comprendían sesiones de desarrollo de contenidos teóricos, como prácticas llevadas a cabo sobre los nodos disponibles. Los participantes han estado en contacto desde el primer momento con los nodos sensores, lo cual potenció aún más el rendimiento de las actividades. El curso se basado en la aplicación de algoritmos de complejidad creciente, los que se esclarecían a medida que se incorporaban nuevos conceptos y normas de programación de los nodos. Al final de cada una de las sesiones se realizaba una puesta en común de las actividades, logros y dificultades, las que se atendían clarificaban en el próximo encuentro. En la tabla 1 se resumen las sesio-

nes planificadas, temas tratados y objetivos perseguidos con el abordaje teórico; así como también las actividades prácticas establecidas.

Tabla 1. - Contenidos desarrollados en Capacitaciones

Sesión 1: Introducción
Teoría: Disparadores de WSN. Principio de funcionamiento de WSN. Aplicaciones. ***Práctica:*** Programación C++, Estructuras básicas para nodos iSense. Herramienta iShell.
Sesión 2: Arquitecturas WSN
T: Nodos WSN. Optimización. Caracterización. Principios de diseño. Redes data céntricas. ***P:*** Despliegues, sinks, intermedios y leaf, Códigos
Sesión 3: Hardware
T: Componentes. Diferencias. Criterios de selección. Factores limitantes. Costos ***P:*** Configuración diferentes módulos
Sesión 4: Aplicaciones y Firmware para nodos
T: Programación, Capacidades de SO. Jerarquía de Clases. ***P:*** Instalación del iSense SDK. Instalaciones en nodo sensor.
Sesión 5: Sistemas Operativos
T: SO en WSN. Funcionalidades básicas. Control de eventos. Soluciones. ***P:*** Ciclos en el S.O. Arranque, eventos y tareas. Temporizadores. Manejo de Memoria.
Sesión 6: Encaminamiento
T: Ruteo. Convergencia. Métodos. Multisalto ***P:*** Algoritmos de enrutamiento, Flooding and hop-based. Tree Routing.
Sesión 7: Sincronización
T: Necesidad. Limitaciones. Interacciones usuario, inter WSN, Mundo real. Desafíos. ***P:*** pruebas de protocolos. Pruebas de algoritmos (LTS, TPSN and HRTS),
Sesión 8: Simulación
T: Fundamentos. 802.15.4a. Simulaciones: factores a considerar. ***P:*** Simuladores en WSN. Instalación y ejecución.

3.3 Actividades de Formación Práctica

Se ha pretendido que la formación práctica representara un distintivo de calidad de los talleres sin descuidar la profundidad y rigurosidad de la fundamentación teórica y la reflexión, como componentes del aprendizaje

Estas tareas en todos los casos han sido ejecutadas con éxito por los participantes. Se listan a continuación algunos de los prototipos actualmente en evolución como parte de proyectos de I+D del grupo de trabajo constituido:

A) Simulación de Redes de Sensores inalámbricos mediante interfaz Web. La simulación por computadora ha permitido a los científicos e ingenieros experimentar fácilmente con ambientes virtuales, alcanzando un nuevo nivel de detalle el análisis de las aplicaciones naturales y artificiales; que fuera desconocido en las primeras etapas del desarrollo científico. Se sabe que modelar analíticamente a las WSN es una tarea complicada, dado que se tiende a realizar análisis simplificados. Toda simulación requiere de un modelo apropiado basado en fundamentos teóricos y sobre todo, de fácil implementación práctica [6], dado que los resultados de la simulación se extrapolan del escenario particular de análisis, con determinadas presunciones, que ciertas veces no encierran al comportamiento real, comprometiendo seriamente con ello la credibilidad de las simulaciones.

En éste proyecto se pretende: a) avanzar en el estado del arte en cuanto simulación WSN, b)Analizar propiedades y eventos necesarios para reproducir el comportamiento de una Red; c) Diseñar la interfaz web de obtención de parámetros, incorporación de los archivos particulares del proyecto y visualización de resultados de simulación; d) Diseñar una solución del lado del servidor Web para procesamiento de datos colectados y generación de resultados de simulación; e) Desarrollar un prototipo de interfaz de simulación de WSN basado en la Web que a través de una interfaz pueda obtener parámetros de simulación, incorporar archivos particulares del proyecto, procesar los datos ingresados y generar los resultados de la simulación; f) Realizar pruebas para comprobar el funcionamiento correcto del sistema.

Se pretende integrar el potencial de la herramienta de simulación Shawn [7] con todas las ventajas de los sistemas basados en la Web. Se plantea incrementar la capacidad del servidor utilizado en las simulaciones por medio de procesamiento distribuido.

B) Sistema Basado en Redes de Sensores Inalámbricos para la Optimización de Recolección de Residuos Domiciliarios en Ciudades Inteligentes. Sabido es que más de la mitad de la población mundial vive hoy en ciudades, y Naciones Unidas estima que el 70% de la población habitarán en centros urbanos en el año 2050. Es primordial por ello, mantener la armonía entre los aspectos espacial, social y ambiental de las localidades, así como entre sus habitantes. En este nuevo escenario sociológico y demográfico, con claros efectos económicos, políticos y medioambientales, cobra fuerza el concepto de ciudad inteligente.

El fin de éste proyecto es diseñar un prototipo de sistema, que utilice los datos generados por WSNs, para determinar que contenedores de residuos urbanos [8] ameritan ser recogidos; calculando con ello una ruta óptima de recolección, pretendiendo resolver problemas de gestión con implementaciones inteligentes.

Aquí se estudian antecedentes sobre Ciudades Inteligentes, Internet del Futuro, y WSN; pretendiendo caracterizar el funcionamiento del sistema de recolección de residuos de la ciudad de Posadas; definir componentes de hardware y software a utilizar en el proyecto, desarrollar software para los nodos sensores, que permita detectar el nivel de llenado de un contenedor; desarrollar un prototipo para captura de datos desde la Red de Sensores Inalámbricos para el cálculo de ruta óptima, y la visualización de la ruta en un mapa; realizar pruebas de laboratorio verificando el funcionamiento completo del sistema.

C) Plataforma para la publicación de datos de Redes de Sensores Inalámbricos, orientada a la visión de la Internet de las Cosas. Se pretende concebir una plataforma para la captura, almacenamiento y publicación de datos de Redes de Sensores Inalámbricos, persiguiendo específicamente: estudiar las tecnologías WSN; evaluar las alternativas existentes para la publicación de datos de WSN útiles para la visión de la IoT; diseñar un prototipo de plataforma que permita la captura, almacenamiento y publicación de los datos obtenidos de la WSN; probar el prototipo en un escenario donde se verifiquen las posibilidades de aplicación práctica como una solución valedera a problemas existentes en diferentes ámbitos.

3.4 Tecnologías Utilizadas

Como plataforma base para los proyectos descriptos se han utilizado equipos con un módulo principal iSense. El hardware iSense se proporciona junto a un firmware operativo y de red modular, permitiendo la generación de aplicaciones pequeñas pero completas; proveyendo una base sólida para el desarrollo rápido de aplicaciones. Brinda una API C++ para el nodo hardware, funcionalidades de sistema operativo y una amplia variedad de protocolos de red.

Ilustración 1. Nodo sensor iSense

En los equipos de desarrollo se utilizó una plataforma PC+Linux en sus distribuciones Ubuntu y Debian. Se instalaron los paquetes make, cmake and gcc++. La plataforma iSense, que incluye al microprocesador Jennic, precisa para el desarrollo de aplicaciones el compilador ba-elf-g++, que asegura la integración perfecta de las librerías.

3.5 Evaluación

Para evaluar el éxito/fracaso de los talleres realizados, se consideraron los siguientes parámetros: Características multidisciplinarias de los participantes capacitados; Cantidad de participantes rechazados por elevada cantidad de postulantes; Calidad de los participantes, expresando el número de participantes con un alto potencial de propagar la capacitación recibida; y la retroalimentación de los participantes.

Los resultados de los diferentes talleres organizados han revelado: interés en el tema WSN; exceso de candidatos a participar en los talleres; buen nivel de los participantes, quienes han aplicado lo aprendido para resolver un problema de la vida real específico.

Como seguimiento de las actividades de formación propuestas, se ha mantenido contacto con todos los participantes, de los cuales el 60% se encuentra actualmente involucrado en diferentes proyectos referidos a las WSN.

Por otra parte, se han identificado como principales debilidades y amenazas al costo de equipos; desconocimiento de tecnología a nivel de decisión y a la infraestructura de soporte en los lugares de implementación.

4 Impactos

4.1 Incorporación de WSN a la Currícula en Carreras de Informática

En el contexto descripto, al observarse el interés de los estudiantes por las competencias prácticas adquiridas en las sesiones de capacitación, y habiendo identificando también una genuina motivación por parte de los alumnos en volcar estas competencias en la producción de sus propias tesinas de grado, la UGD decidió incorporar al nuevo plan de estudios de su carrera Ingeniería en Informática en el área de Tecnologías Aplicadas, una nueva asignatura electiva referida a la Tecnología WSN e Internet de las Cosas.

La decisión tiene que ver con la evolución que se propone actualmente a nivel internacional desde la Joint Task Force on Computing Curricula (ACM & IEEE-CS) [9], como también en los ámbitos de debate sostenidos en nuestro país promovidos por la RedUNCI y la RIISIC para la estandarización de contenidos básicos e intensidad de formación práctica de las carreras en Informática.

La asignatura tiene relación directa con Redes y Comunicaciones siguiendo un enfoque bottom-up, construyendo la comprensión de las redes desde sus niveles más bajos o físicos hacia los más altos en los modelos de referencia, con una modalidad organizativa del tipo taller, basada en el Aprendizaje Basado en Problemas (ABP) [10]. Las metas a alcanzar son: Comprender y aplicar conocimientos inherentes a las WSN, en función de los desafíos presentados por la tecnología, a saber: Eficiencia Energética, Capacidad limitada de procesamiento, ancho de banda y almacenamiento, altos niveles de error, escalabilidad y robustez.

Por otra parte, los contenidos sintéticos definidos incluyen: Motivación para el estudio de las WSN. Sensores. Génesis de WSN. Desafíos. Aplicaciones. Arquitectura de nodos. Sistemas Operativos. Modelo de Referencia. Capa física. Control de acceso al medio. Capa de red. Gestión de energía. Sincronización. Localización. Seguridad y Programación en WSN.

En todos los casos se prioriza la realización de actividades de resolución de problemas abiertos, de proyecto y diseño, en la búsqueda de inculcar un espíritu de investigación tendiente a producir descubrimientos que permitan nuevos desarrollos, articulándose perfectamente con el Taller de Tesis de Grado/Trabajo Final de carrera vigente en la universidad.

4.2 Radicación de Nuevos Proyectos de I+D

El desarrollo de los talleres de WSN ha consolidado la conformación de un grupo de investigación y desarrollo, integrado por docentes e investigadores de la UNaM y UGD, habiéndose planteado varias (6) tesinas de grado en el contexto [11], incluyendo algunas con relación directa con la industria [12] .

La UGD ha asignado recursos para que varios alumnos en etapa de preparación de su tesis de grado sean asimilados como becarios en el mencionado proyecto, recursos estos que son solventados con fondos propios de la universidad, logrando con ello involucrar a estudiantes de grado en proyectos de investigación, desarrollo e innovación tecnológica con beneficios reales y concretos para su formación, exponiendo a los estudiantes a interesantes proyectos en su contexto regional socio-productivo y al proceso reflexivo crítico involucrado en la investigación científico-tecnológica.

5 Conclusiones y Trabajos Futuros

Se considera que los talleres de implementación de WSN descriptos en este trabajo han resultado exitosos, ya que han promovido la conformación de un grupo de I+D, en cuyo contexto se desarrollan diferentes tesinas de grado y proyectos fomentado el desarrollo local de soluciones que buscan resolver problemas del entorno regional. Los talleres han resultado la fuente de actualización curricular de una de las carreras en Informática que ofrecen las universidades involucradas en la experiencia reportada.

El objetivo último ha sido siempre conducir a cada uno de los participantes a la obtención de una solución para la problemática particular propia, descartando en todo momento las actividades conductoras a un único tipo determinado de proyecto, tendiendo a la co-creación del conocimiento [13]. En este mismo sentido, se ha desarrollado una Wiki, con documentación construida de forma cooperativa en los talleres de capacitación ofrecidos y en la que se proveen además guías prácticas para la instalación del entorno de desarrollo SDK de iSense y la instalación del simulador Shawn.

Finalmente, es destacable mencionar que en los próximos talleres a desarrollar se prevé enfocarse en IPv6 y seguridad en WSN, con la certeza que la demanda de este tipo de capacitación es creciente, estando persuadidos acerca de que la red IPng será soporte indefectible de todas las redes de sensores, alentando el advenimiento de Internet de las cosas.

Bibliografía

[1] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.,* vol. 3, no. 3, 1999.

[2] ITU, "ITU Internet Reports 2005: The Internet of Things," 2005.

[3] D. Dickson, "Buenas y malas noticias sobre la 'brecha científica'," SciDev.Net, 2009. [Online]. Available: http://tiny.cc/00jtww.

[4] MINCyT, "PLAN NACIONAL DE CIENCIA,TECNOLOGÍA E INNOVACIÓN: Lineamientos estratégicos 2012-2015," Buenos Aires, 2012.

[5] E. Sosa, "Tesis Doctoral: Contribuciones al Establecimiento de una Red Global de Sensores Inalambricos Interconectados," Universidad de La Plata, 2011.

[6] E. Egea-López and e. al., "Simulation Tools for Wireless Sensor Networks," in

Summer Simulation Multiconference - SPECTS 2005, 2005.

[7] S. Fekete, A. Kroller, S. Fischer and D. Pfisterer, "Shawn: The fast, highly customizable sensor network simulator," Braunschweig, 2007.

[8] S. Longhi, D. Marzioni, E. Alidori, G. Di Buo, M. Prist, M. Grisostomi and M. Pirro, "Solid Waste Management Architecture Using Wireless Sensor Network Technology," in *5th International Conference on New Technologies, Mobility and Security (NTMS)*, 2012.

[9] The Joint Task Force on Computing Curricula, "Computer Science Curricula 2013," Association for Computing Machinery & IEEE-Computer Society, February 2013.

[10] L. V. Morales P., "Aprendizaje basado en problemas," *Theoria,* vol. 13, no. 1, pp. 145-157, 2004.

[11] E. Sosa, D. Godoy, R. Neis, G. Motta, R. Luft, D. Sosa, H. Bareiro and P. Quiñones, "Internet del Futuro y Ciudades Inteligentes," in *XV Workshop de Investigadores en Ciencias de la Computación 2013*, Paraná, 2013.

[12] A. Quiñones, D. Godoy and E. Sosa, "Redes Inalámbricas de Sensores: Una experiencia en la Industria del Té," in *5º Congreso Argentino de AgroInformática(en prensa)*, Córdoba, 2013.

[13] B. Regeer and J. Bunders, "Knowledge co-creation: Interaction between science and society. A Transdisciplinary Approach to Complex Societal Issues," Advisory Council for Research on Spatial Planning, Nature and the Environment/Consultat, La Haya, 2009.

El desarrollo de la comprensión lectora en las carreras de Informática

Sonia Rueda

Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur

Abstract. El desarrollo de las **competencias comunicativas** involucra entre otras capacidades la **comprensión lectora**. En las carreras de Informática las materias iniciales demandan cierto nivel de desarrollo de esta capacidad pero también brindan la oportunidad de reforzarla. En estas asignaturas se utiliza lenguaje natural y distintos lenguajes artificiales. Este trabajo describe los diferentes modelos de textos y de lenguajes a través de los cuáles se fortalece la comprensión lectora, en cada una de las materias de Programación de las carreras ofrecidas por el Departamento de Ciencias e Ingeniería de la Computación de la UNS.

Keywords: Diseño curricular basado en competencias. Comprensión Lectora. Lenguaje natural y lenguaje artificial.

1 Introduction

La formación de un profesional en Informática requiere, entre otros aspectos, el desarrollo del razonamiento lógico, la argumentación, la organización de la información y la apropiación del lenguaje de la ciencia y la tecnología. Todas estas capacidades demandan a su vez de *competencias comunicativas* que permitan elaborar e interpretar *actos del habla*.

Cada nivel educativo presupone cierto nivel de desarrollo en las competencias comunicativas y asume la responsabilidad de reforzarlo. Sin embargo, en el ámbito universitario es poco frecuente que se establezca con precisión lo que se asume adquirido, lo que se espera desarrollar y las acciones y actividades concretas que se realizan para alcanzar este desarrollo.

La *comprensión lectora* es una capacidad básica fundamental para la competencia comunicativa de un alumno que comienza a cursar una carrera en el nivel superior. Es también imprescindible para un graduado de una carrera universitaria. Ahora bien, ¿qué nivel de comprensión lectora se espera de un ingresante? ¿cuál debería ser el nivel de desarrollo de esta capacidad en un graduado?

El objetivo de este trabajo es analizar y describir los diferentes *modelos de textos*, *elementos discursivos* y *lenguajes* con los cuales se fortalece la comprensión lectora en las materias de Programación de las carreras ofrecidas por el Departamento de Ciencias e Ingeniería de la Computación (DCIC) de la Universidad Nacional del Sur (UNS). Los programas de estas materias incluyen a los contenidos básicos

establecidos para el trayecto Algoritmos y Lenguajes en la resolución ME 786/09. La siguiente sección presenta el marco conceptual del trabajo. A continuación se describe brevemente la metodología de enseñanza adoptada en el las materias de Programación. Luego se vincula esta metodología con el desarrollo de la comprensión lectora, indicando los modelos de textos y lenguajes que se utilizan en estas asignaturas. Por último, se ofrecen algunas conclusiones y lineamientos para el trabajo futuro.

2 Marco Conceptual

Sobre finales de los ´90 comienza en Europa la *Reforma Universitaria* a partir de una declaración conjunta que dio inicio a un *proceso de convergencia educativa*. Los principales objetivos de este proceso, que continúa vigente, son promover la cooperación para garantizar la calidad educativa, facilitar la movilidad de alumnos, docentes e investigadores y establecer un sistema internacional de créditos que permita comparar titulaciones [1]. Dada la diversidad de los sistemas educativos europeos, la expectativa no es uniformar los planes de estudio, sino adoptar un *modelo curricular basado en competencias*. El modelo busca disminuir la brecha entre la formación académica acreditada y las oportunidades para obtener un puesto de trabajo; propone establecer lo que un graduado debe *saber*, *saber hacer* y *saber ser* para poder insertarse eficazmente en el sistema productivo.

Este enfoque tiene antecedentes previos al proceso de convergencia europeo, pero es a partir de esta declaración que ha tomado relevancia en el ámbito universitario. Desde entonces el término *competencia* ha sido definido de maneras diferentes [2]. Algunos autores lo consideran equivalente a *capacidad*. Otros le asignan al término capacidad un significado estático, vinculado a una aptitud potencial, mientras que una competencia sería una capacidad puesta en acción, esto es, aplicada a una situación concreta de modo que puede evaluarse la eficacia. Bajo otro criterio el concepto de capacidad se refiere a la integración de conocimientos, aptitudes y habilidades que se adquieren en el ámbito formativo, mientras que las competencias son capacidades aplicadas al desempeño profesional.

En este trabajo definimos **competencia** como "*un conjunto de conocimientos, capacidades y actitudes que se articulan y aplican eficazmente para resolver una situación o problema concreto en un contexto académico o laboral*". Definimos **capacidad** como una "*aptitud o habilidad que se desarrolla a partir de la integración y acumulación de aprendizajes significativos"*.

Las competencias requeridas en el sistema productivo coinciden parcialmente con las demandadas en el ámbito universitario. En las empresas esperan que los profesionales de Informática sean competitivos, proactivos, asuman riesgos, puedan liderar equipos de trabajo, enfrenten desafíos, etc. Claramente no todos los puestos de trabajo de una empresa demandan estas características, pero el crecimiento de un profesional con frecuencia está tan ligado a estas cualidades, como a su formación académica. En general los contenidos, metodologías de enseñanza y mecanismos de evaluación del plan de estudios de una carrera, no están específicamente orientados al desarrollo o valoración de estas cualidades.

Las diferencias entre las expectativas de uno y otro ámbito y la responsabilidad de la Universidad con relación al desarrollo de competencias laborales, ha sido tema de

debate en la última década. Más allá de estas diferencias, la habilidad social, el trabajo en equipo y en particular la *competencia comunicativa* son reconocidas y valoradas tanto en la etapa formativa, como durante el desempeño profesional.
Consideramos que la **competencia comunicativa** es *un conjunto de conocimientos, capacidades y actitudes articulados y aplicados eficazmente para comprender y elaborar actos del habla*. Un **acto del habla** es una *acción* que transmite un *mensaje* desde un *emisor* hacia un *receptor* o *destinatario* en un *lenguaje* compartido. Esta competencia requiere reconocer no solo el significado explícito o literal de un mensaje, *lo que se dice,* sino también las implicaciones, el sentido explícito o intencional, *lo que el emisor quiere decir* o *lo que el destinatario quiere entender*.
Este trabajo se refiere específicamente en la *comprensión lectora*, considerando que se trata de una de las capacidades fundamentales para la competencia comunicativa, tanto en el ámbito académico como profesional. Restringimos el análisis a las materias de Programación de las carreras ofrecidas por el DCIC.

2.1 La comprensión lectora

La lectura es un proceso de interacción entre el pensamiento y el lenguaje. Este proceso requiere reconocer los símbolos del lenguaje y las reglas que estructuran el uso de estos símbolos, pero también *comprender* el significado de lo leído. La comprensión permite elaborar un significado para el texto. El lector elabora un significado a partir de conceptos que a su vez tienen significado para él y de sus **estrategias para la comprensión lectora** [3].
Este trabajo se concentra en la comprensión de dos **tipos de textos**: *expositivos* y *descriptivos*. Estos tipos de texto aparecen en diferentes **modelos** como*: apuntes de cátedra, trabajos prácticos, libros, manuales, tutoriales, instructivos, artículos científicos, páginas web, presentaciones de diapositivas, etc.* El formato puede ser impreso o digital. Este último puede incluir recursos como sonido, animación e hipervínculos que permiten una lectura no lineal.
Cada modelo de texto puede incluir diferentes **modos discursivos**, como *definiciones*, *comparaciones*, *enumeraciones*, *ejemplificaciones*, *síntesis*, etc. Cada modelo de texto y cada modo discursivo se distingue por su *forma* y sobre todo por su *intencionalidad*.
El **nivel de comprensión** depende de la complejidad del texto y también de la formació en el tema y la *intencionalidad* del lector respecto a la profundidad de la lectura. Una *lectura primaria* o *literal* puede ser suficiente para reconocer entidades, conceptos, sucesos, acciones y relaciones explícitas. La *lectura inferencial* permite identificar causas-efectos y explicitar información o asociaciones, que están señaladas implícitamente. El reconocimiento de metáforas y analogías implica un nivel comprensión aun mayor. También se requiere un nivel de lectura comprensiva más elevado cuando una actividad exige emitir juicios o comparar entidades fundamentando la valoración. La *lectura crítica* tiene entonces un carácter evaluativo donde interviene el criterio del lector y su conocimiento acerca de lo leído [8].

2.2 La metodología de enseñanza basada en la resolución de problemas

Desde la década del ’80 se han desarrollado metodologías de enseñanza basadas en la resolución de problemas con enfoques diferentes:

- Aprender ***desde*** la resolución de problemas implica abordar los contenidos conceptuales de una asignatura específica a partir de problemas.
- Aprender ***sobre*** la resolución de problemas se refiere a aprender acerca de los procesos y estrategias cognoscitivas que aplicamos cuando resolvemos problemas, para controlarlos en forma consciente y deliberada.
- Aprender ***a*** resolver problemas se orienta a que los alumnos adquieran destreza en la resolución de problemas generales, reconociendo la importancia de cada etapa del proceso de resolución.

Estos enfoques no son alternativos sino que pueden complementarse y son consistentes con la práctica reflexiva, el trabajo en equipo y por proyectos, el aprendizaje autónomo y comprometido. Demandan docentes capaces de trabajar con grupos heterogéneos, generar y gestionar situaciones de aprendizaje, entre otras competencias docentes. Las situaciones de aprendizaje deben constituir un desafío, pero al mismo tiempo deben ser una meta alcanzable [6][7].

Las metodologías de enseñanza basadas en la resolución de problemas resultan particularmente valiosas para el diseño curricular basado en competencias, dado que la resolución de problemas es una competencia fundamental tanto en el ámbito académico como laboral. Es también una metodología útil para el desarrollo de la competencia comunicativa ya que la interpretación del problema demanda capacidad para la comprensión lectora y por lo general la construcción de la solución exige habilidades para la producción escrita.

3. Las materias de Programación de las carreras del DCIC

Este trabajo analiza y describe la capacidad de comprensión lectora que se asume alcanzada y que se aspira a desarrollar en cuatro materias del área Programación: *Resolución de Problemas y Algoritmos (RPA)*, *Introducción a la Programación Orientada a Objetos (IPOO)*, *Estructuras de Datos (EdD)* y *Tecnología de Programación (TdP)*. Estas asignaturas forman parte de los planes de estudio de las carreras ofrecidas por DCIC de la UNS.

Todas adoptan una metodología de enseñanza basada en la resolución de problemas. Bajo esta concepción, el desarrollo de un *programa* puede pensarse como la construcción de una solución para un problema real, la solución es un modelo escrito en un lenguaje de programación **[4]**.

Antes de comenzar a cursar RPA, los alumnos deben haber aprobado el curso de nivelación *Análisis y Comprensión de Problemas,* uno de cuyos principales objetivos es justamente reforzar la comprensión lectora aplicada específicamente a la resolución de problemas **[5].** Durante este curso se refuerzan las estrategias para la comprensión lectora, en particular la inferencia. Se plantean problemas que permiten reconocer la importancia de interpretar adecuadamente un enunciado para identificar la incógnita, los datos explícitos e implícitos y las restricciones. Los enunciados y las soluciones incluyen textos en distintos lenguajes gráficos.

En RPA se asume que los alumnos pueden interpretar adecuadamente el enunciado de un problema como parte del proceso de resolución, son capaces de aplicar *estrategias* básicas para la comprensión lectora de distintos *modelos de textos* y reconocen diferentes *modos discursivos*. Al completar el cursado de TdP la expectativa es que

hayan reforzado las estrategias de lectura sobre un conjunto más amplio de modelos de textos, puedan comprender diferentes lenguajes artificiales y decidir el nivel de lectura que se requiere para comprender un texto, considerando la intencionalidad.

4 Lenguajes naturales y artificiales

Un **lenguaje natural** es un lenguaje con propósito general, generado espontáneamente y usado por humanos para comunicarse. Aunque existen reglas sintácticas que restringen la forma de combinar los símbolos del lenguaje, todo lenguaje natural admite cierto nivel de ambigüedad y permite que algunas expresiones tengan un significado literal diferente al sentido figurado.

En la secuencia de materias de Programación es importante que los docentes planifiquen la progresión de aprendizajes [7], consideren y decidan premeditadamente cuál es el nivel de lectura que va a requerir cada texto, qué modos discursivos y modelos de textos se proponen y diseñen actividades considerando las habilidades que asumen adquiridas para la compresión lectora y las que aspiran profundizar, en cada etapa de la secuencia.

Un **lenguaje artificial** es una notación con un propósito específico y puede estar definido formalmente, en este caso tiene una sintaxis estricta y una semántica precisa. Como otras ciencias, la Matemática posee un lenguaje artificial que simplifica y clarifica la comunicación, designando de una manera exacta sus contenidos a través de términos y símbolos específicos. La comunicación se simplifica y clarifica si los alumnos pueden interpretar estos términos y símbolos, como así también el modo de estructurarlos. Si por el contrario, desconocen el lenguaje utilizado, la comprensión del contenido se dificulta, por más exacta y precisa que sea una expresión. De hecho, una de las dificultades de los alumnos que ingresan a una carrera universitaria se relaciona justamente con la interpretaciones de textos escritos en una *notación formal* como la que brinda la Matemática.

Las materias de Programación ofrecen múltiples oportunidades de aplicar notaciones formales, en particular el lenguaje Matemático aprendido en el Nivel Polimodal y profundizado en las asignaturas de Ciencias Básicas. Es importante que se reflexione explícitamente acerca de los beneficios de utilizar una notación rigurosa y se generen actividades que incluyan entre sus objetivos la "lectura" de textos escritos usando esta notación. Una especificación como *Sea S una secuencia de la forma* $s_1\ s_2\ \ldots\ s_{m-1}\ s_m$ *tal que* $s_i \leq s_{i+1}$ *con* $1 \leq i < m$.va a requerir que el docente describa la notación al comenzar el cursado de RPA y ayude a interpretarla tanto en las clases teóricas como prácticas, manteniendo por supuesto la uniformidad en el lenguaje. Al completar esta asignatura la lectura de este tipo de expresiones debería ser fluida.

Así, la *semántica de las operaciones* que los alumnos deben implementar comienza a describirse combinando lenguaje natural y una notación formal desde la primera materia de Programación. La definición recursiva de factorial, el algoritmo para computar el máximo común divisor entre dos números naturales, las propiedades de secuencias de elementos, la especificación formal de las operaciones del tipo de dato abstracto Pila, son algunos de los ejemplos con los cuáles se aprenden a usar notaciones formales para establecer la semántica de las operaciones.

La interpretación de cada una de estas especificaciones se puede complementar con algunos ejemplos significativos, que luego pueden ser usados como casos de prueba en la verificación. Es importante que los alumnos entiendan que la solución tiene que ser general, no solo funcionar correctamente para los casos de prueba específicos.

4.1 Lenguaje natural

En las cuatro materias de Programación en las clases teóricas se utilizan presentaciones de diapositivas. Cada docente aplica diferentes criterios y estilos para armar sus presentaciones y combinarlas con otros recursos, como por ejemplo el pizarrón, el ambiente de programación o animaciones para ilustrar la ejecución de algunos programas. Cualquiera sea el criterio y estilo, las presentaciones siempre incluyen y destacan el *vocabulario técnico* que se va introduciendo junto con los contenidos conceptuales propios de cada materia. Así, el lenguaje natural conocido por los alumnos, aumenta con nuevas palabras, propias de la disciplina.

Aunque asumimos que los alumnos pueden reconocer diferentes *modos discursivos*, en general resulta valioso detenerse para resaltar las características de cada uno de ellos con respecto a la *forma* y sobre todo la *intencionalidad*. Asimismo es pertinente remarcar antes de una evaluación, que cuando se pida la definición de un concepto, no se considerará válida una respuesta que sólo aporte algunos ejemplos, aunque se refieran a ese concepto; del mismo modo, si una consigna requiere una comparación entre dos recursos, es probable que dos descripciones independientes no se consideren válidas como respuesta, excepto que ambas describan a los conceptos en términos de los mismos atributos y la comparación sea evidente, aunque no esté explícita.

La lectura de manuales y tutoriales es una actividad en la cual los alumnos tienen que ir adquiriendo destreza progresivamente. Es importante que se incluyan actividades que requieran interpretar textos de temas no desarrollados en clase, pero que al mismo tiempo tengan una complejidad adecuada para ser comprendidos de manera autónoma. Cuando los manuales o tutoriales están disponibles en versión digital es posible estimular la lectura no lineal proponiendo actividades que requieran navegar a través de hipervínculos.

En RPA los contenidos conceptuales y el lenguaje de programación utilizado se presentan en detalle en las clases teóricas a través de la resolución de problemas. El lenguaje natural aparece entonces en las descripciones, definiciones, comparaciones y enunciados. Los enunciados incluidos en trabajos prácticos y evaluaciones, con frecuencia se complementan con textos escritos en lenguajes artificiales, por ejemplo algoritmos, expresiones matemáticas, diagramas.

Es importante que el docente reflexione cuando produce una actividad, cuál va a ser nivel de lectura que demanda. Más allá de la dificultad que pueda entrañar el problema, si los datos, la incógnita y las restricciones están formulados explícitamente, la comprensión de texto no demandará mayor esfuerzo. Si en cambio la interpretación del enunciado requiere un proceso de abstracción para reconocer los datos relevantes o un proceso de inferencia que permita identificar datos que están implícitos, la capacidad de comprensión toma un rol más importante para la resolución.

En las dos primeras materias de Programación, como así también en el curso de nivelación, se incluyen problemas equivalentes en complejidad respecto a las estrategias de resolución, pero con diferentes requerimientos respecto al nivel de

comprensión lectora. La expectativa es justamente identificar las dificultades para poder superarlas.

En IPOO los elementos básicos del lenguaje de programación se describen brevemente en las clases prácticas y se asume cierta autonomía por parte de los alumnos para profundizar los temas presentados y evacuar dudas consultando manuales, tutoriales y por supuesto a los docentes. Algunos paquetes y clases provistas por el lenguaje se describen parcialmente y se propone como actividad investigar acerca de otros los servicios. En ningún momento la expectativa es que los alumnos los memoricen. Los recursos provistos por el lenguaje para soportar los conceptos centrales de la programación orientada a objetos, se presentan en teoría, se aplican en la práctica y se refuerzan a través de la lectura de bibliografía.

En EdD y TdP la lectura profunda de la bibliografía propuesta es indispensable para las evaluaciones finales y la realización de proyectos. Una de las cuestiones que afecta a la adaptación de los alumnos al ámbito universitario es justamente el nivel de profundidad con el que leen y nivel de lectura con que se espera que lean. Una lectura superficial de un texto puede ser muy adecuada entre una clase y otra, de modo que el docente pueda profundizar en un tema o incluso presentar uno nuevo, asumiendo que los alumnos comprendieron las nociones básicas del contenido presentado previamente. El mismo nivel de lectura no va a ser adecuado cuando se espera que comprendan en profundidad un tema, por ejemplo para aplicar los contenidos para la resolución de un problema o la realización de un proyecto.

En cada materia es importante estimula la reflexión acerca de cuándo es suficiente realizar una lectura *primaria* y cuándo se espera que alcancen un nivel de comprensión más profundo a través de una lectura *inferencial* o más aun, *crítica*. Un aspecto a remarcar es que la lectura de un resumen de un texto puede ser más compleja que la del texto original, porque en el resumen suelen quedar implícitas asociaciones y se eliminan aclaraciones, observaciones, analogías y ejemplos que favorecen la comprensión. Este hecho no siempre es evidente para los alumnos y con ejemplos adecuados puede ilustrarse claramente. Este tipo de actividades, que en principio no son parte de los objetivos de estas asignaturas, pueden resultar fundamentales para favorecer capacidades como la comprensión lectora.

En cada una de las tres primeras materias de Programación se utiliza un entorno de desarrollo diferente. Desde el punto de vista de la comprensión lectora un ambiente de desarrollo demanda interpretar las opciones provistas por los distintos tipos de menúes, los mensajes de error del compilador y la información que nos brinda el depurador. Actualmente los alumnos están habituados a interactuar con diferentes interfaces gráficas de usuarios y en general se adaptan rápidamente a explorar y utilizar una aplicación de software. También es común que la instalación de una aplicación no implique una dificultad. Aun así, es conveniente proponer alguna actividad que implique leer un instructivo como por ejemplo para instalar un entorno de desarrollo, crear un paquete o un ejecutable.

4.2 Lenguajes Artificiales

Uno de los objetivos de las materias de Programación es que los alumnos interpreten textos escritos y construyan soluciones usando diferentes lenguajes. Algunos de estos lenguajes tienen una especificación formal como los *lenguajes de programación, lenguajes de modelado, lenguajes gráficos de especificación sintáctica, notaciones*

matemáticas, etc. Otros lenguajes artificiales no tienen reglas formales que los definan como por ejemplo los *lenguajes de diseño de algoritmos* y los *lenguajes para modelar la evolución de la memoria.*

En cada materia es necesario planificar la presentación de los lenguajes en las clases teóricas respecto a cómo se utilizan y combinan en los enunciados de problemas y proyectos en las clases prácticas y en las evaluaciones parciales y globales. Un objetivo más importante aun, es que los alumnos *aprendan a aprender* lenguajes artificiales. El enfoque basado en resolución de problemas es fundamental en este sentido porque cada lenguaje se aprende *usándolo*. Así, más allá de describir los aspectos sintácticos y semánticos, es necesario centrarse especialmente de las cuestiones pragmáticas de cada notación. El uso de distintas notaciones artificiales contribuye a desarrollar la *autonomía en el aprendizaje*. Es muy importante articular la secuencia de acuerdo a la cual se presentan y utilizan cada una de estas notaciones en cada asignatura, para que los alumnos puedan desarrollar gradualmente la capacidad de comprender y producir textos escritos utilizando distintos lenguajes.

En RPA se propone un **lenguaje de diseño de algoritmos** que, con una sintaxis menos rigurosa que el lenguaje de programación, favorece la construcción de un modelo para la resolución de cada problema planteado. El objetivo de esta etapa es identificar las estructuras de control requeridas para la resolución. El lenguaje de diseño resulta útil cuando el problema puede ser dividido en subproblemas o la resolución implica combinar al menos dos estructuras de control. Si el problema es muy simple, no se destaca la ventaja de diseñar un algoritmo antes de comenzar a escribir código en el lenguaje de programación.

También en RPA se introduce un **lenguaje para modelar la evolución de memoria**, que permite construir diagramas que muestran una traza de la ejecución de un programa o subprograma. Aunque existen diferentes formas de construir estos diagramas, es importante acordar una notación uniforme dentro de una misma asignatura. En IPOO se usan **diagramas de evolución de referencias** que modelan la memoria, pero en este caso en el paradigma orientado a objetos.

Los **lenguajes gráficos de especificación sintáctica** se utilizan en las cuatro materias descriptas en este trabajo. La sintaxis de cada mecanismo provisto por los lenguajes de programación usados, se especifica a través de *diagramas sintácticos***.** En RPA la aparición de un diagrama requiere detenerse en explicar la notación, más allá del objeto mismo que se está describiendo. En general no se destina una clase específica a describirla, sino que se presenta en forma transversal a otros contenidos. Es muy importante articular las materias y acordar en qué momento se espera que los alumnos tengan autonomía para la lectura de este tipo de diagrama.

En IPOO los alumnos interpretan diagramas sintácticos, reconocimiento especificaciones condicionales, iterativas y recursivas. La "lectura" de un diagrama sintáctico permite reconocer instrucciones y expresiones válidas, distinguir aspectos sintácticos y semánticos de un lenguaje de programación y comprender la necesidad de que un lenguaje de programación tenga una sintaxis rigurosa.

Los diagramas sintácticos se utilizan también para especificar la forma de una secuencia de elementos que deben ser procesados. La resolución del problema requiere interpretar la especificación y luego diseñar el algoritmo para procesar la secuencia e implementarlo en un lenguaje de programación. Si los alumnos están

familiarizados con este tipo de diagrama, la formalización es una ayuda porque permite descubrir las estructuras de control adecuadas y facilita el diseño del algoritmo. En caso contrario, la "lectura" del diagrama puede representar una dificultad adicional al problema propiamente dicho.

En IPOO se introducen los elementos básicos de un **lenguaje de modelado** para construir *diagramas de clases* y en las materias que siguen se presentan otro tipo de recursos para elaborar distintos tipos de diagramas. En IPOO muchos de los enunciados incluyen un diagrama de clases en donde se especifican las decisiones de diseño de cada clase y las relaciones que las vinculan. La interpretación del diagrama, incluyendo las responsabilidades de cada clase, es una de las capacidades que se aspira desarrollar en esta asignatura. En las materias que siguen no solo se interpretan modelos escritos en UML sino que muchos problemas requieren producir diseños usando este lenguaje.

La comprensión lectora de textos producidos usando los lenguajes artificiales mencionados, es una capacidad que las materias de las áreas Ingeniería de Software y Sistemas, asumen adquiridas previamente. El desarrollo de esta capacidad se alcanza como tanto en las materias de Programación como en las asignaturas del área Fundamentos de Ciencias de la Computación.

Conclusiones y trabajo futuro

Las competencias comunicativas son reconocidas y valoradas tanto en el ámbito académico como en el sistema productivo. Los planes de estudio de las carreras ofrecidas por el DCIC de la UNS no incluyen asignaturas orientadas específicamente al desarrollo de la comprensión lectora, la producción escrita y la oralidad. Estas capacidades se consideran transversales, de modo que el diseño curricular de las asignaturas se realiza asumiendo cierto nivel adquirido de cada una de ellas y se compromete a reforzarlas.

En las materias del área Programación de estas carreras se adoptó hace más de 20 años una metodología de enseñanza basada en la resolución de problemas. El objetivo principal de aprendizaje está vinculado evidentemente con el desarrollo de programas, para lo cual se presentan varios lenguajes de programación.

La comprensión del lenguaje natural se enriquece con la lectura de diferentes *modelos de textos*. Es importante que los docentes hagan notar a los alumnos las particularidades de cada modelo, especialmente en lo que se refiere a la comunicación, e indiquen también el nivel de lectura que se espera que realicen de cada uno, en cada momento. La caracterización de tres *niveles de lectura* puede ser suficiente en este sentido: primaria, inferencial y crítica.

Cualquiera sea el modelos de texto puede incluir diferentes *modos discursivos*. Estos modos son conocidos previamente por los alumnos pero la reflexión explícita acerca de las características y sobre todo la intencionalidad de cada uno, contribuye a fortalecer las competencias comunicativas. En particular, las definiciones aumentan el vocabulario incorporando términos propios de la disciplina y enriqueciendo el lenguaje natural.

La introducción progresiva de diferentes *lenguajes artificiales*, con distintos grados de rigurosidad sintáctica y semántica, contribuye a desarrollar la competencia

comunicativa. Es importante articular adecuadamente la presentación y aplicación de cada nueva notación, de modo que cada una constituya una ayuda y no un obstáculo. Algunos lenguajes van a ser descriptos en detalle, siempre aplicados a situaciones concretas, en otros se va a requerir mayor autonomía de aprendizaje por parte de los alumnos. En cada caso debe reflexionarse y establecerse explícitamente las competencias previas que se asumen adquiridas.

Este trabajo se concentra específicamente en la comprensión lectora abordada desde cuatro asignaturas de una misma área. Una extensión natural es considerar la producción escrita y la oralidad. El trabajo puede extenderse también para abordar el desarrollo de la competencia comunicativa en otras áreas de la disciplina. Una vez que se haya avanzado en estos dos sentidos la expectativa es analizar otras competencias reconocidas tanto en el sistema educativo como productivo.

Referencias

[1] Declaración de Bolonia. Declaración conjunta de los Ministros Europeos de Educación. Bolonia (1999) http://www.ond.vlaanderen.be/hogeronderwijs/bologna/

[2] Mastache, Anahí: *Formar personas competentes. Desarrollo de competencias tecnológicas y psicosociales.* Novedades Educativas.ISBN:978-987-538-199-5 (2007).

[3] Castro Fox Guillermina, González Nora, Monti Carla, Palmucci Daniela: *Estrategias para la Lectura Comprensiva. Un abordaje al discurso científico-pedagógico.* Universidad Nacional del Sur. (2005)

[4] Rueda Sonia, García Alejandro: *Análisis y Comprensión de Problemas: Fundamentos, Problemas Resueltos y Problemas Propuestos*. Notas del curso de nivelación. Universidad Nacional del Sur. Argentina Págs.: 90. (2004)

[5] Rueda Sonia, Señas Perla: *Un enfoque basado en la resolución de problemas para la enseñanza de la Programación Orientada a Objetos* Bahía Blanca III Congreso en Tecnología en Educación & Educación en Tecnología (Teyet) (2008)

[6] Perrenoud Philippe: *Diez nuevas competencias para enseñar*. Editorial Graó. ISBN 978-84-7827-321-8 (2004)

[7] Whimbey Arthur, Lochhead Jack, Narode Ronald: *Problem Solving & Comprehension* Wearset Ltd. Boldon ISBN 978-0-415-50221-4 (2013)

[8] Alonso Tapia, Jesús: Leer, comprender y pensar: Desarrollo de estrategias y técnicas de evaluación. MEyC. CIDE. Madrid. ISBN: 84-369-2270-0 (1992)

Extensión del Lenguaje y Modelo Simplesem con Soporte para Paralelismo

Lucas L. Diez de Medina, Gustavo Wolfmann y Orlando Micolini

Facultad de Ciencias Exactas, Físicas y Naturales. Universidad Nacional de Córdoba
Av. Velez Sarsfield 1611 - Córdoba – Argentina
lucaslt89@gmail.com
gwolfmann@gmail.com
omicolini@compuar.com

Resumen El modelo de ejecución planteado por Carlo Ghezzi y Mehdi Jazayeri, conocido como Simplesem, no contempla lenguajes con instrucciones de ejecución paralela. El objectivo es extender esta herramienta educativa incorporando al modelo la existencia de múltiples procesadores primitivas de lenguaje que permitan expresar conceptos básicos de paralelismo. Para ello, se desarrolló una herramienta que permite analizar de manera gráfica y sencilla el impacto de programas multihilo sobre instrucciones de bajo nivel, que operan directamente sobre la memoria compartida de una máquina virtual. Esta extensión permite representar la semántica operacional de lenguajes paralelos que requieren procesadores multihilo, comunes en la actualidad.

1. Introducción

Las dificultades a las que se enfrenta un alumno al estudiar el proceso que atraviesa un programa escrito en un lenguaje de alto nivel, hasta poder ser ejecutado por el procesador. Principalmente esto se debe a dos motivos: La complejidad de los lenguajes de programación de alto nivel existentes, y la del hardware sobre el cual se ejecutan.

En su libro "Programming Language Concepts", Carlo Ghezzi y Mehdi Jazayeri [1] propusieron una alternativa para solucionar estos problemas. En primer lugar, propusieron un modelo de ejecución que consiste en una máquina formada por un procesador simple, una memoria de código y una memoria de datos. El procesador es capaz de ejecutar cuatro instrucciones elementales que conforman las primitivas del modelo Simplesem. En segundo lugar, definieron una serie de lenguajes de programación de alto nivel, de complejidad creciente, que coinciden con diversas categorizaciones de lenguajes. A partir de estas dos propuestas, lograron simplificar el proceso de aprendizaje de la relación entre los lenguajes de programación, y los recursos de la computadora.

Las memorias de datos y de código de la máquina Simplesem están formadas por celdas. En la memoria de código, las celdas almacenan instrucciones, y se utiliza un puntero para seleccionar la celda que contiene la siguiente instrucción a ejecutar. En la memoria de datos se almacena un valor por cada celda. Las

operaciones se realizan directamente sobre la memoria de datos (no hay registros intermedios). Existen algunas celdas en las que se almacenan valores específicos utilizados para el funcionamiento de la máquina.

Con respecto a las instrucciones, el modelo Simplesem consta de tres instrucciones compuestas (**set**, **jump** y **jumpt**), y una instrucción especial para indicar la finalización del programa (**halt**). Por instrucciones compuestas se entiende aquellas instrucciones que pueden contener expresiones, y el resultado de dichas expresiones será lo que se utilice como operandos de la instrucción.

Las primitivas del modelo Simplesem permiten implementar la semántica operacional de un conjunto de lenguajes de alto nivel. Ghezzi y Jazayeri definieron lenguajes de alto nivel para demostrar el funcionamiento del modelo Simplesem.

El primer lenguaje, conocido como C1, sólo posee tipos de datos simples, e instrucciones simples (no tiene soporte para funciones). Sólo pueden utilizarse tipos de datos que permitan determinar los requerimientos de memoria de manera estática, como enteros, arreglos de tamaño fijo y estructuras. El segundo lenguaje, conocido como C2, incorpora la posibilidad declarar rutinas en el programa, que pueden contar a su vez con variables locales. Las rutinas no soportan llamadas recursivas, no pueden recibir parámetros, ni devolver valores de retorno. El tercer lenguaje, C3, agrega la posibilidad de que las rutinas devuelvan un valor, y pueden ser llamadas recursivamente.

El lenguaje de alto nivel utilizado en este trabajo es una extensión del lenguaje C3, en la cual se incorporan primitivas de ejecución paralela, y se demuestra cómo puede implementarse la semántica operacional de este nuevo lenguaje, con primitivas del modelo Simplesem.

Contribución: El presente trabajo se basó en extender el modelo de Ghezzi y Jazayeri (que sólo estudiaba lenguajes monohilo), para abarcar también lenguajes con primitivas de ejecución paralela. La principal motivación fue la predominancia de los procesadores multicore, y la importancia que ha adquirido la programación concurrente actualmente. La propuesta consiste en la definición de un nuevo lenguaje de alto nivel, y la extensión del conjunto de instrucciones Simplesem para implementar la semántica operacional de dicho lenguaje.

La ejecución paso a paso de un programa en un modelo multiprocesador es compleja, por lo que la heramienta cuenta con una interfaz gráfica de la máquina Simplesem, que permite seguir y controlar la ejecución de de las instrucciones de manera interactiva.

El proceso de desarrollo constó de tres etapas. La primera consistió en la extensión del modelo básico monohilo, a un modelo multihilo. En la segunda etapa se definieron las instrucciones básicas simplesem que permiten la ejecución de programas paralelos. Finalmente, se definió la gramática de un lenguaje de alto nivel, con primitivas de paralelismo, cuya semántica operacional puede ser implementada con la extensión del modelo Simplesem.

Importancia: Con la herramienta lograda al finalizar este proyecto, se ampliará el campo de estudio del modelo Simplesem a lenguajes multihilo de ejecución paralela, de manera tal que aprovechando la simplicidad del modelo, podrán estudiarse conceptos más complejos inherentes a la programación concurrente.

2. Desarrollo e Implementación

La utilización del modelo Simplesem en el ámbito educativo ha demostrado tener buenos resultados en el proceso de aprendizaje de los lenguajes de programación de alto nivel. Sin embargo, dicha propuesta se limita sólo a lenguajes secuenciales, con un único hilo de ejecución. Proponemos ahora una extensión del modelo para aplicarlo a la enseñanza de lenguajes multihilo.

2.1. Procesador multihilo

La primera extensión se realiza sobre el diseño del procesador de la máquina Simplesem. Podrán crearse hasta cuatro hilos paralelos. Las memorias de código de todos los hilos contendrán las mismas instrucciones, por lo que el comportamiento específico deberá determinarse de acuerdo al identificador de cada hilo, de manera tal que cada uno ejecute sólo las instrucciones que le corresponden. La memoria de datos será compartida por todos los hilos. Cada hilo tendrá su propio puntero a instrucción.

2.2. Nuevas Instrucciones Simplesem

Para realizar una extensión a lenguajes paralelos, se agregaron nuevas primitivas al modelo Simplesem.

processors n. Esta instrucción provoca que se creen n hilos, cada uno con su propio IP, y con una memoria de código independiente del resto. Al momento de su creación, todos los hilos estarán formados por las mismas instrucciones, y será el programador quien deba dotar de un comportamiento distinto a cada hilo.

barrier. Cada vez que el procesador deba ejecutar la instrucción barrier de un hilo, deberá verificar si todos los otros hilos están en la misma instrucción. De no ser así, la ejecución deberá bloquearse hasta que todos los hilos estén en la misma instrucción.

wait n. esta instrucción es utilizada para lograr la implementación de semáforos a través de instrucciones Simplesem. Cada vez que una instrucción wait se ejecuta sobre la celda n de la memoria de datos, si el valor almacenado en la celda n es 0 la ejecución del hilo se bloquea. De no ser así, se decrementa el valor de n en 1, y se continua con la ejecución.

Esto podría hacerse a través de múltiples, instrucciones Simplesem, pero la modificación de los permisos de un semáforo debe ser una operación atómica, por lo que se necesita una única instrucción para tal fin.

numHilo. El indicador numHilo se utiliza como palabra clave dentro de las instrucciones Simplesem para referirnos al hilo que actualmente se está ejecutando. Cada hilo posee un identificador (de 0 a 3) que se utilizará para generar un offset de una posición en la memoria de datos, para acceder a los datos que corresponden a cada hilo.

2.3. Lenguaje C3P – Extensión del Lenguaje C3

El enfoque abordado consiste en utilizar una variable paralela, que será distinta para cada hilo que se ejecute. El resto de las variables, tanto globales como locales, son compartidas por todos los hilos.

A continuación se presentan las incorporaciones que se hicieron al lenguaje C3, para generar un nuevo lenguaje al que llamamos C3P, con primitivas de ejecución en paralelo.

Sentencia par_for. Suponiendo que i es la variable paralela (distinta en todos los hilos), la estructura de esta sentencia es:

```
par_for 4 (i = 0; i < 8; i++){
  suma = suma + i;
}
end_par_for;
```

El número *4* representa la cantidad de hilos que quieren crearse. El límite del ciclo for (en este caso 8) debe ser un número divisible por N (4), de manera tal que cada hilo procesará *8 / 4 = 2* valores de i y los sumará al resultado final. Las instrucciones que se encuentren dentro del bloque *par_for* serán ejecutadas por los cuatro hilos.

Sentencia barrier. Dentro de un bloque *par_for*, necesitamos un mecanismo para sincronizar todos los hilos. Cada vez que un hilo encuentra una sentencia *barrier*, este se bloquea hasta que todos los otros hilos lleguen a la misma sentencia. En ese momento todos los hilos se desbloquean y pueden continuar su ejecución.

Si colocamos una sentencia *barrier* debajo de la sentencia de suma del ejemplo anterior, todos los hilos terminarán de procesar el primer valor de i antes de poder procesar el segundo valor.

Sentencia wait. Para abordar problemas típicos de concurrencia (en los cuales se utilizan múltiples hilos), debemos contar con construcciones como semáforos o monitores. Para ello, se incorporó la sentencia *wait* al lenguaje C3P.

Esta sentencia se ejecuta sobre una variable, por ejemplo *wait(i)*. Si el valor de la variable es mayor que cero, se disminuye en uno su valor y se incrementa en uno el contador de programa del hilo que ejecutó la instrucción wait. Si el valor de la variable es 0, el contador de programa no se modifica.

Sentencia notify. La sintaxis de esta sentencia es *notify(variable)*. Fue incorporada debido a que habitualmente se utiliza para incrementar en uno los permisos de un semáforo. El efecto de la sentencia *notify(i)* es el mismo que el de la sentencia *i++*. Si algún hilo se encontraba bloqueado por una instrucción *wait* sobre la variable *i*, al haberse incrementado el valor de esta última, dicho hilo se desbloqueará.

Funciones con parámetros. Otra de las funcionalidades con las que no contaba el lenguaje C3, eran funciones que recibieran parámetros.

El lenguaje C3P incorpora la posibilidad de declarar funciones que reciban múltiples parámetros. Los parámetros se guardan en el registro de activación al llamar a una función, antes de las variables locales de la misma. Los parámetros pueden utilizarse dentro de una función como si fueran variables locales.

Restricciones del Lenguaje C3P: Para simplificar el proceso de compilación del lenguaje C3P, se realizaron dos simplificaciones respecto al lenguaje C3. En primer lugar, todas las variables son de tipo entero, ya que soportar otros tipos de datos implicaba que el compilador realizara una detección del tipo de datos, para realizar la reserva de memoria en función de esto.

La segunda simplificación fue con respecto a las funciones. El lenguaje C3P sólo soporta la declaración y utilización de una única función. Esto se debe a que incorporar más de una función implica realizar un manejo de la tabla de símbolos del compilador, para conocer la ubicación en memoria de las instrucciones y el tamaño del registro de activación de cada función.

2.4. Implementación

La herramienta desarrollada se llevó a cabo en tres etapas: un proceso de formalización del lenguaje C3P, la implementación del compilador y la implementación de la máquina de ejecución abstracta (intérprete) de las instrucciones Simplesem.

2.5. Formalización del Lenguaje C3P

La primera etapa consiste en extender el lenguaje de alto nivel agregando sentencias a la gramática del lenguaje C3. Las nuevas reglas de producción del lenguaje C3P se muestran a continuación[1]

[1] La gramática completa del lenguaje C3P puede encontrarse en[3]

```
statement : assign ";"
          | "read" "(" iden ")" ";"
          | "write" "(" expr ")" ";"
          | iterate
          | condition
          | "return" expr ";"
          | "wait" "(" iden ")" ";"
          | "notify" "(" iden ")" ";"
          | call ";"
          | parallel_statement
          | comment

parallel_statement : parallel_for
                   | "barrier" ";"

parallel_for : parallel_for_beginning statement_block parallel_for_end

parallel_for_beginning : "par_for" numbers for_definition

parallel_for_end : "end_par_for" ";"
```

2.6. Implementación del Compilador

Para la implementación del compilador, se utilizó lenguaje PERL, junto con el módulo Parse::RecDescent [4], que permite implementar un parser descendente recursivo. Este módulo provee una sintaxis para describir formalmente la gramática de un lenguaje de programación. El compilador tiene como objetivo generar las instrucciones Simplesem que implementen la semántica operacional de un programa escrito en lenguaje C3P. Para tal fin, hace uso de un parser descendente recursivo que analiza el código fuente escrito en lenguaje C3P, genera elementos representables por instrucciones Simplesem, y finalmente obtiene dichas instrucciones.

El parser recibe como entrada un programa escrito en lenguaje C3P y la gramática del lenguaje.

A medida que se van encontrando coincidencias del código fuente con las reglas de producción del lenguaje C3P, se van generando las instrucciones Simplesem correspondientes.

Cada producción permite incorporar *acciones*, en las cuales tendremos acceso a los valores encontrados (es decir a los elementos de la sentencia que coincidió con la regla en cuestión). Con el uso de estas *acciones*, se generan las instrucciones Simplesem en función de cada sentencia específica.

2.7. Implementación del Intérprete

Para la elaboración del intérprete se utilizó lenguaje C++, junto con la biblioteca de clases Qt.

El intérprete consiste de un editor de texto integrado, donde puede escribirse código fuente, compilarlo y obtener las instrucciones Simplesem en una tabla que representa la memoria de instrucciones y la memoria de datos.

Cuenta con controles que permiten realizar distintos modos de ejecución. Podemos procesar una instrucción de un hilo específico, una instrucción de todos los hilos a la vez, o realizar una ejecución continua hasta que el intérprete encuentre una instrucción halt.

La Figura 1 muestra cómo se ve la interfaz gráfica del intérprete.

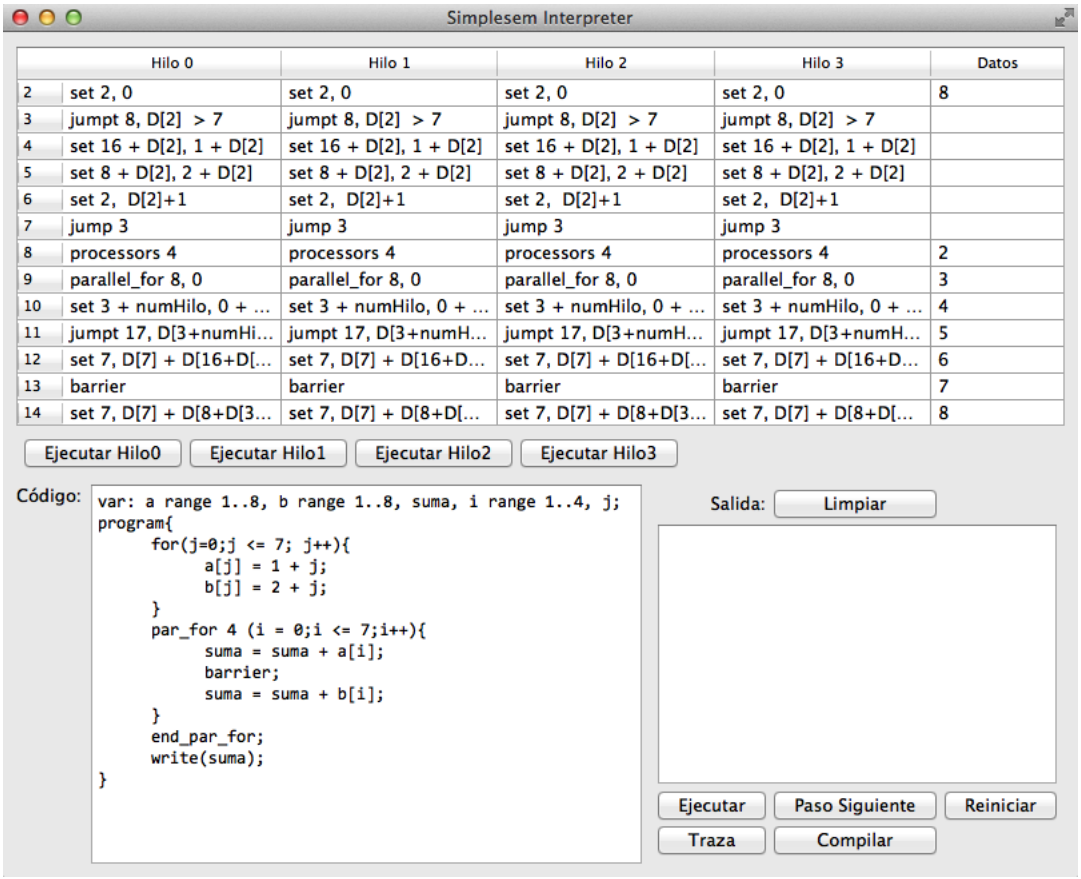


Figura 1. Interprete Simplesem con cuatro hilos ejecutándose en paralelo

3. Uso de la herramienta

El intérprete cuenta con 6 ejemplos integrados que utilizan todas las primitivas del lenguaje C3P. Dichos ejemplos pueden ser cargados, modificados y compilados a instrucciones Simplesem desde la misma herramienta. Tanto el código

fuente como versiones distribuidas para Windows, Linux y Mac OS pueden ser descargados desde el repositorio de la herramienta[2].

El siguiente ejemplo muestra la implementación del problema de los filósofos (Dijkstra, 1965) en lenguaje C3P, para cuatro filósofos, cada uno de los cuales está representado por un hilo.

```
var: tenedor range 1..4, filosofo range 1..4, j;
program{
    //Inicializacion de los semaforos (tenedores).
    for(j=0;j <= 3; j++){
        tenedor[j] = 1;
    }
    par_for 4 (filosofo = 0;filosofo <= 3; filosofo++){
        while(1 > 0){
            if(filosofo == 3){
                write(filosofo*10+1); //Pensando
                wait(tenedor[0]); //Toma los recursos
                wait(tenedor[3]);
                write(filosofo*10+0); //Comiendo
                notify(tenedor[3]); //Libera los recursos
                notify(tenedor[0]);
            }
            else{
                write(filosofo*10+1); //Pensando
                wait(tenedor[filosofo]); //Toma los recursos
                wait(tenedor[filosofo+1]);
                write(filosofo*10+0); //Comiendo
                notify(tenedor[filosofo+1]); //Libera los recursos
                notify(tenedor[filosofo]);
            }
        }
    }
    end_par_for;
}
```

Las instrucciones Simplesem del bloque paralelo de este ejemplo se muestran en la Figura 2.

[2] https://github.com/lucaslt89/SimplesemExtension

	Hilo 0	Hilo 1	Hilo 2	Hilo 3	Datos
7	processors 4	processors 4	processors 4	processors 4	0
8	parallel_for 4, 0	parallel_for 4, 0	parallel_for 4, 0	parallel_for 4, 0	0
9	set 3 + numHilo, 0 + numH...	set 3 + numHilo, 0 + numH...	set 3 + numHilo, 0 + numH...	set 3 + numHilo, 0 + numH...	0
10	jumpt 29, D[3+numHilo] >...	jumpt 29, D[3+numHilo] >...	jumpt 29, D[3+numHilo] >...	jumpt 29, D[3+numHilo] >...	1
11	jumpt 27, 1 <= 0	jumpt 27, 1 <= 0	jumpt 27, 1 <= 0	jumpt 27, 1 <= 0	
12	jumpt 20, D[3+numHilo] !...	jumpt 20, D[3+numHilo] !...	jumpt 20, D[3+numHilo] !...	jumpt 20, D[3+numHilo] !...	
13	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	
14	wait 7+0	wait 7+0	wait 7+0	wait 7+0	
15	wait 7+3	wait 7+3	wait 7+3	wait 7+3	
16	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	
17	set 7+3, D[7+3]+1	set 7+3, D[7+3]+1	set 7+3, D[7+3]+1	set 7+3, D[7+3]+1	
18	set 7+0, D[7+0]+1	set 7+0, D[7+0]+1	set 7+0, D[7+0]+1	set 7+0, D[7+0]+1	
19	jump 26	jump 26	jump 26	jump 26	
20	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	
21	wait 7+D[3+numHilo]	wait 7+D[3+numHilo]	wait 7+D[3+numHilo]	wait 7+D[3+numHilo]	
22	wait 7+D[3+numHilo] + 1	wait 7+D[3+numHilo] + 1	wait 7+D[3+numHilo] + 1	wait 7+D[3+numHilo] + 1	
23	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	set write, D[3+numHilo] * 1...	
24	set 7+D[3+numHilo] + 1, ...	set 7+D[3+numHilo] + 1, ...	set 7+D[3+numHilo] + 1, ...	set 7+D[3+numHilo] + 1, ...	
25	set 7+D[3+numHilo], D[7...	set 7+D[3+numHilo], D[7...	set 7+D[3+numHilo], D[7...	set 7+D[3+numHilo], D[7...	
26	jump 11	jump 11	jump 11	jump 11	
27	set 3 + numHilo, D[3+num...	set 3 + numHilo, D[3+num...	set 3 + numHilo, D[3+num...	set 3 + numHilo, D[3+num...	
28	jump 10	jump 10	jump 10	jump 10	
29	end_par_for	end_par_for	end_par_for	end_par_for	
30	halt	halt	halt	halt	
31					

Figura 2. Instrucciones Simplesem del bloque paralelo en el problema de los filósofos.

En el estado actual de ejecución, el filósofo 0 posee los dos recursos (tenedores); el filósofo 1 se encuentra bloqueado esperando a que se libere un recurso; el filósofo 2 tomó un recurso, y puede tomar el siguiente sin bloquearse; y el filósofo 3 está bloqueado esperando un recurso que adquirió el filósofo 1. Si bien todos los hilos poseen las mismas instrucciones, los hilos 0, 1 y 2 ejecutarán las instrucciones entre las celdas 20 y 25, y el hilo 3 las instrucciones entre las celdas 13 y 18. Esto se debe a que tres filósofos deben tomar los recursos en el mismo orden, y el cuarto los deberá tomar en orden inverso para evitar interbloqueos.

4. Conclusión

El modelo de ejecución presentado por Carlo Ghezzi y Mehdi Jazayeri brindó una herramienta de gran riqueza educativa, simplificando la enseñanza y el aprendizaje de conceptos fundamentales de los lenguajes de programación.

Con el trabajo aquí presentado, se logró extender el modelo para incorporar lenguajes paralelos, además de brindarse una herramienta que engloba de manera práctica, todos los conceptos planteados teóricamente, desde la definición de un lenguaje hasta la ejecución de un programa escrito en dicho lenguaje sobre una máquina Simplesem.

El lenguaje de alto nivel planteado, el parser/compilador desarrollado, y la interfaz gráfica del modelo de ejecución, conforman una herramienta completa, que permite estudiar todas las etapas por las que atraviesa un programa escrito en un determinado lenguaje, desde que comienza a compilarse, hasta que finaliza su ejecución.

La ejecución paso a paso de un programa paralelo en una interfaz gráfica, a través de la cual se puede interactuar con la máquina Simplesem, ayuda a la comprensión de conceptos como el interbloqueo o la inhanición. Permite además analizar el impacto del acceso concurrente a la memoria de datos, y la manera en que los hilos de un programa paralelo deberán sincronizarse para ejecutar un programa de la manera esperada.

En este trabajo, no se realizó un estudio sobre los beneficios a nivel educativo que la herramienta brinda, dejando esta tarea sujeta a una futura investigación.

Si bien existen muchas mejoras posibles a la extensión planteada, consideramos que el trabajo realizado es de gran utilidad en el ámbito educativo, ya que brinda una manera práctica de abordar el estudio de los lenguajes de programación y la relación que estos tienen con los recursos de la computadora.

Referencias

1. Carlo Ghezzi y Mehdi Jazayeri, Programming Language Concepts, Tercera edición, 1996. Publicado el 23 de junio de 1997 por John Wiley & Sons. 448 Páginas. ISBN: 0471104264.
2. Sitio web oficial de Perl: www.perl.org Consultado en diciembre de 2012.
3. Lucas Leandro Diez de Medina Quintar. Extensión del Modelo Simplesem a un Lenguaje Paralelo. Tesis de Grado disponible en http://goo.gl/9QwTJp – Publicada en Julio de 2013.
4. Parse::RecDescent Documentation. Disponible en: http://search.cpan.org/perldoc?Parse::RecDescent Consultado en diciembre de 2012.

Conformando repositorios de datos de la comunidad educativa en la Universidad Nacional de La Plata Un caso de estudio.

Javier Díaz[1], María Alejandra Osorio[2], Ana Paola Amadeo[3]

Laboratorio de Investigación en Nuevas Tecnologías Informáticas, Facultad de Informática,

Universidad Nacional de La Plata

Calle 50 y 120, 1900 La Plata, Buenos Aires. Argentina

[1]jdiaz@unlp.edu.ar, [2]aosorio@cespi.unlp.edu.ar, [3]pamadeo@linti.unlp.edu.ar

Abstract. Este artículo presenta las iniciativas llevadas a cabo por el CeSPI, el Centro Superior para el Procesamiento de la Información de la Universidad Nacional La Plata, respecto a análisis de información e integración de sistemas desarrollados ad-hoc para atender la complejidad y diversidad de las distintas realidades de la UNLP, o a través de implementaciones propuestas por el Ministerio de Educación de la Nación para la gestión universitaria. La construcción de repositorios consolidados de datos de alumnos, docentes y no docentes ha presentando nuevos desafíos relacionados con la confiabilidad de los datos y mecanismos de actualización, que han fomentado la revisión de procedimientos, el análisis de herramientas de big data y han traído aparejado beneficios en distintos niveles. Las redes sociales y los intentos de integración y análisis sobre ellas también son contemplados en este artículo.

Palabras Clave: análisis de información, integración de servicios, redes sociales, SAML, UNLP, Argentina

1 Introducción

La Universidad Nacional de La Plata es una institución de educación superior pública, 3° en el país en cantidad de alumnos, según el Anuario Estadístico del año 2010.Está ubicada en la ciudad de La Plata, capital de la provincia de Buenos Aires, Argentina. Desde su fundación en 1905, es pionera en estudios y desarrollos culturales, artísticos y científicos de avanzada La docencia, la investigación y la extensión configuran los pilares básicos de esta Universidad. Agrupa a 17 Facultades de las ramas más diversas del saber, donde estudian más de 100 mil alumnos. En los últimos años se registra un promedio de inscripciones cercano a los 23.000 aspirantes, de los cuales se trasforman en alumnos alrededor de 18.500. De sus aulas egresan anualmente alrededor de 5.800 estudiantes. La oferta académica de la UNLP incluye 111 carreras de grado -157 títulos- y 170 de posgrado (el 85% están acreditadas o en trámite, por la Comisión

Nacional de Evaluación y Acreditación Universitaria –CONEAU-), además de unos 500 cursos de posgrado. Además cuenta con 49 cátedras libres dependientes de la Presidencia, que se suman a las muchas que funcionan en las Facultades. En el pregrado, la oferta académica incluye cinco Colegios Preuniversitarios con una matrícula cercana a los 5 mil alumnos [1]

El Centro Superior para el Procesamiento de la Información, de aquí en adelante CeSPI, es el centro de servicios informáticos de la UNLP. Su misión es *Propiciar el uso y apropiación de las Tecnologías de la Información y Comunicación y los cambios sociales necesarios para su aprovechamiento, que contribuyan a mejorar las funciones de educación, investigación científica y tecnológica y extensión universitaria que desarrolla la Universidad Nacional de La Plata; aportando a una sociedad sostenible social y ambientalmente* [2] Creado en 1959, su función es colocar a la tecnología al servicio de la Institución. En el Centro se realizan las tareas relacionadas con los distintos sistemas que brindan servicios a la Universidad. Estos sistemas comprenden la liquidación de sueldos de los empleados, el manejo curricular de los alumnos de las respectivas unidades académicas y la tarea que sostiene éstas actividades: la administración y el soporte técnico de la red de datos, los servicios de Internet y la propia infraestructura del Centro. En el año 2009 los procesos de *Gestión de Requerimientos de Servicios e Información de Sistemas Académicos, y los Servicios de Auditoría y Consultoría Tecnológica*, tuvieron reconocimiento internacional al certificar la norma de calidad ISO 9001:2008. Las certificaciones fueron concedidas por la organización especializada TÜV Rheinland Argentina S.A[3] dependiente de la alemana TÜV Rheinland Group tras haber verificado que el diseño y la implementación del sistema de gestión es el adecuado para la ejecución de los procesos y cumple con las exigencias de la norma. En el año 2010 y 2011 se han superado tanto las auditorías internas a cargo de un profesional calificado, como las externas de certificación y seguimiento por parte del organismo TÜV, con resultados altamente satisfactorios. En el año 2012, el alcance de la certificación comprendió a los procesos de *Gestión de Requerimientos de Sistemas Académicos, de Seguridad de la Información, de Minería, Análisis de Datos y de Servicios de Auditoría y Consultoría Tecnológica.* Cabe recordar que los procesos de Gestión de Requerimientos de Servicios e Información de Sistemas Académicos son los que actualmente utilizan todas las Unidades Académicas de la Universidad Nacional de La Plata.

En el CeSPI se han desarrollado aplicaciones en respuesta a las crecientes demandas de una institución pública compleja, con un volumen importante de usuarios que generan numerosos requerimientos. Además de los alumnos, que según el último informe de indicadores del año 2012[4] suman 108.934 personas, se incluyen 12.056 docentes y 2.969 personal administrativo. En este contexto, es función del CeSPI garantizar la interconexión a través de las redes de datos hasta las aplicaciones que se utilizan para la operatoria diaria de la Universidad. En la figura 1 se presentan las distintas aplicaciones que da soporte actualmente, algunos por directivas del Ministerio de Educación y otros desarrollos propios. De estos últimos, la mayoría utilizan un mecanismo de autenticación Single Sign On implementado a través de SAML, que se detalla en el apartado 3 de éste artículo. Por otra parte, a través de la Dirección de Sistemas Académicos se trabaja en la implantación de software de código abierto como la plataforma virtual Moodle, el sistema de bibliotecas Koha-

Meran así como también de otras soluciones implementadas por el Ministerio de Educación de la Nación, como el sistema de gestión académica SIU Guaraní y SIU Mapuche.

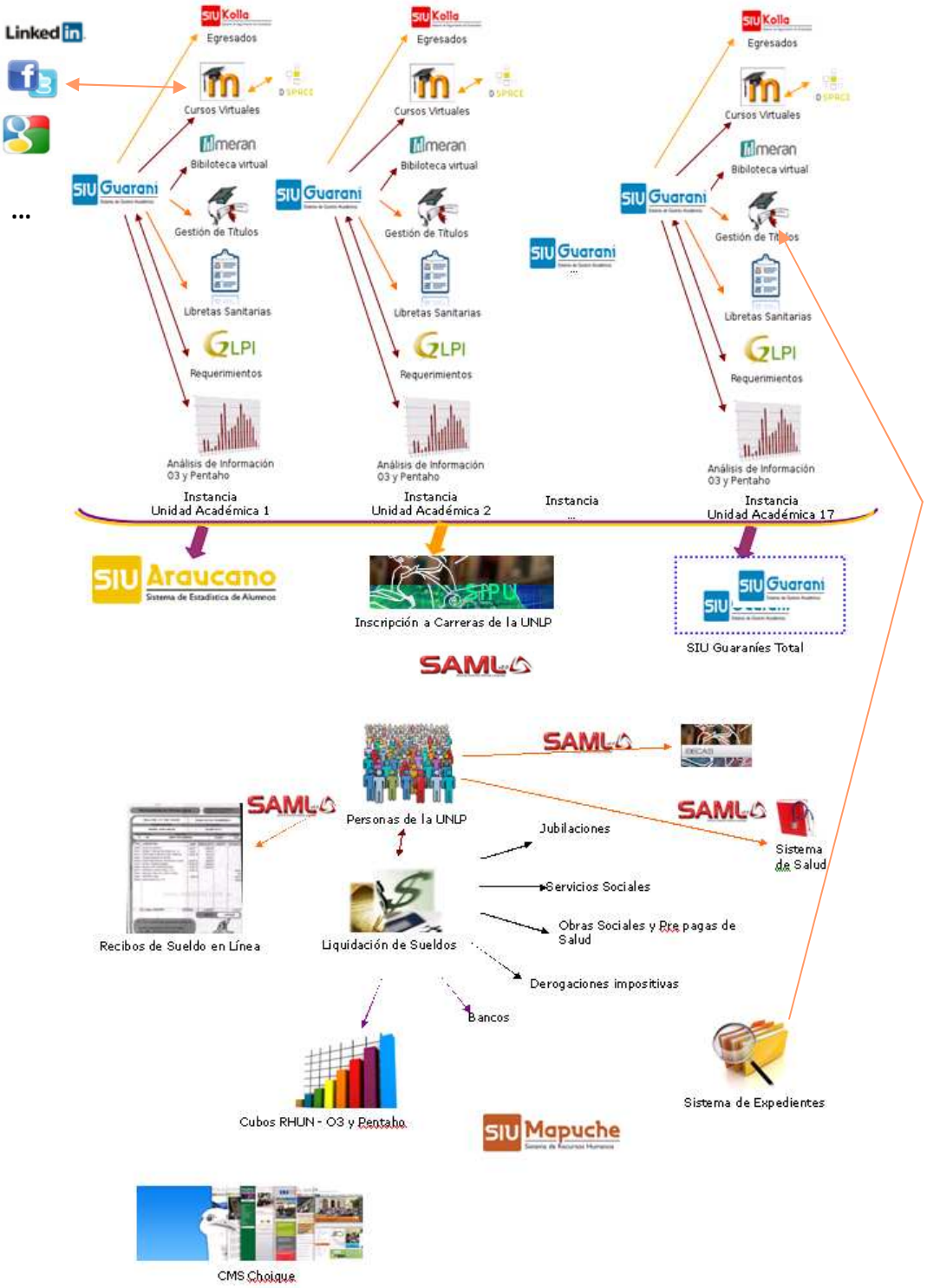


Fig. 1 – Integración de sistemas, desarrollados y/o mantenidos por el CeSPI, utilizando web services (flechas naranjas) o intercambio de archivos (flechas violetas). Todos estos sistemas generan información que se consolida en distintos repositorios: Guaraní Total para alumnos y Personas UNLP para alumnos, docentes y no docentes. Se relacionan también con las redes sociales en un límite difuso que se está desvaneciendo.

La interrelación entre todos los sistemas facilita y promueve iniciativas para el análisis de la información en forma integrada, confiable y concreta para la toma de decisiones.

2 Análisis de la Información

La Dirección de Sistemas Académicos del CeSPI incluye el área de Análisis de la Información. Surge a partir de la creciente demanda de analistas de información que permitan generar conocimiento a partir de los sistemas de gestión operativa que se utilizan hoy en día en toda organización, con distintos fines.
Como mencionamos previamente, esta área ha certificado ISO 9001:2008 para la gestión de requerimientos de los usuarios en el último año. La ampliación del alcance de la certificación para incluir *Minería y Análisis de Datos* obedeció a la creciente demanda por parte de las entidades de gestión de las facultades y el propio Rectorado, de datos confiables, íntegros y consolidados para la toma de decisiones fundamentadas en cada unidad académica y la implementación de planes y políticas que afectan directamente a la comunidad educativa de estudiantes, docentes y personal administrativo. Además, la información recolectada por el área de Análisis de Información de Sistemas Académicos del CeSPI realiza informes periódicos al Ministerio de Educación a través del sistema SIU Araucano.
Desde sus inicios, el CeSPI ha sido el responsable de almacenar y resguardar en forma adecuada la información gestionada por los sistemas de gestión operativa de la universidad, siendo responsable de los reportes emitidos sobre los mismos a partir de solicitudes realizadas por agentes internos o externos, como el Ministerio de Educación de la Nación o el Poder Judicial para verificación de datos.
La evolución de los distintos sistemas de información, la diversificación de los mismos, personas usuarias de distintas aplicaciones y las necesidades de información, por parte de los altos mandos directivos que permitan realizar análisis multidimensionales para la toma de decisiones, ha hecho imprescindible trabajar en un repositorio consolidado de datos del mismo sujeto, el estudiante, almacenados en los sistemas de gestión operativa de uso diario en la universidad.
En particular, en la UNLP se utiliza el sistema SIU Guaraní desde el año 2002, en 15 de las 18 facultades, en 3 direcciones de postgrado y capacitación no docente. El SIU Guaraní es un sistema desarrollado por Consorcio SIU de Universidades, que desarrolla soluciones informáticas para el Sistema Universitario Nacional y organismo del gobierno. Su objetivo es colaborar, a través de los sistemas de información confiables y seguros, con el mejoramiento continuo de la gestión: optimizar los procesos, la calidad de los datos y facilitar la toma de decisiones contando con una sólida base de información. [5] El SIU Guaraní es una de las soluciones ofrecidas para gestionar todas las actividades académicas de las universidades nacionales, desde que los alumnos aspiran a formar parte de la universidad hasta que egresan con su diploma, contemplando la complejidad y heterogeneidad del sistema universitario nacional [6]. Actualmente se encuentra implementado en más de 200 unidades académicas de todo el país.
La implementación del Guaraní implica un proceso de depuración datos analizando los repositorios académicos históricos (dependiendo de las Facultades tienen en línea información de los últimos 20 y hasta 50 años). Durante los meses de febrero y marzo

se registran las actividades más intensivas, a través de las inscripciones a los cursos de dictado regular del ciclo lectivo. Durante los meses de febrero y marzo de este año a través del sistema se registraron más de 311318 inscripciones.

El SIU Guaraní se integra en forma natural con otros sistemas desarrollados por el SIU como el SIU Araucano, SIU Data Warehouse y SIU Kolla. Además de estas soluciones propias, en distintas unidades académicas se utiliza el sistema de código abierto Moodle [7] para la gestión de cursos virtuales, en general como complemento de la actividad presencial. Este sistema está integrado con el sistema SIU Guaraní a través de una interfaz común, que facilita la gestión de las inscripciones a trabajos prácticos entre ambos sistemas. El sistema Moodle se utiliza en la Facultad de Informática desde el año 2005, incluyendo más de 310 cursos y más de 14000 usuarios. Puede consultarse a través de http://catedras.info.unlp.edu.ar, http://cursos.linti.unlp.edu.ar y http://postgrado.linti.unlp.edu.ar La información almacenada en este sistema involucra entregas de tareas, participación en los foros, acceso a los diferentes recursos y demás registros los cuales, al cruzarlos con el desenvolvimiento académico pueden resultar en datos significativos para la toma de decisiones de los directivos de la Facultad. Otras unidades académicas como la Facultad de Ingeniería, de Ciencias Veterinarias, de Ciencias Económicas y Ciencias Médicas también utilizan Moodle como plataforma virtual en sus cursos presenciales.

Por otra parte, para la gestión de la biblioteca se utiliza el sistema de software libre Meran[8], desarrollado por el equipo de desarrollo del CeSPI a partir de Koha [9], el sistema de SL para la gestión de bibliotecas implementado en distintas facultades de la UNLP a partir del año 2003. Este sistema permite gestionar los procesos bibliotecarios y los servicios a los usuarios, como estantes virtuales para las cátedras, la posibilidad de votar un libro o dar recomendaciones. Todo esto integrado también al sistema SIU Guaraní.

La UNLP y muchas de sus unidades académicas están haciendo uso de las redes sociales para poder llegar a los distintos sectores de la sociedad que están vinculados con ella y de esta manera realizar una mejor difusión de las actividades que se llevan a cabo. En el caso de la UNLP, dispone de un perfil en Facebook, Universidad.Nacional.La.Plata, con más de 5800 seguidores. La Facultad de Informática posee también cuentas en FB y Twitter para difusión académica como @InformaticaUNLP con más de 1100 seguidores, @infounlp con más de 500 seguidores entre otras. Además, ciertas cátedras utilizan estos canales para comunicarse con sus estudiantes, como la cátedra de Algoritmos y Estructuras de Datos @ayed_fi, Introducción a los Sistemas Operativos @iso_info_unlp con más de 100 seguidores, Sistemas Operativos, entre otros.

Como se puede observar, se cuentan con distintas fuentes de información relacionadas con el mismo sujeto, el estudiante, que consolidadas en único repositorio permitirá realizar análisis transversales, que sirvan de insumo a grupos interdisciplinarios con distintos perfiles. Esta realidad hace imprescindible aplicar distintas metodologías para la constitución de un Data Warehouse que soporte las distintas estrategias y técnicas para obtener conocimiento a partir de los datos almacenados en distintos sistemas. La construcción de un DW involucra una serie de actividades relacionadas con distintas técnicas para la Extracción, Transformación y Carga soportadas por distintas herramientas como Kettle de Pentaho BI[10], Spago BI[11], Talend BI[12], entre otras. La naturaleza del negocio determina los requerimientos de los usuarios

que en este caso están relacionados con rendimiento académico, comportamiento en entornos virtuales y redes sociales.

Respecto al primer punto, rendimiento académico, se está trabajando en los cubos de análisis de información provistos por el SIU. La información base se toma del sistema SIU Guaraní. En el inicio de la implementación de los cubos de análisis de información, se utilizaba la herramienta O3[13]. Los datos eran tomados directamente de las bases de datos productivas en los momentos en que se registraba menor actividad, en general los fines de semana. A medida que el tamaño de las bases de datos aumentaba y el sistema se comenzó a utilizar en un régimen de 7x24 a través de las interfaces Web, se hizo necesario contar con una instancia replicada, imagen de las bases de datos productivas al primer sábado de cada mes, para generar los cubos de análisis estadísticos y reportes ad-hoc. Para reportes concretos en línea, se continuaba tomando los datos de las BD productivas.

La integración con otros sistemas y la necesidad del desarrollo de nuevos data marts que integren los datos de todas las unidades académicas hizo imprescindible la construcción de un Data Warehouse. Para esto se utilizó la metodología propuesta por Kimball[14] basada en la construcción de pequeños data marts específicos para las distintas áreas de la organización. Es así como hoy se cuenta con un repositorio de información académica que centraliza los datos de todos los alumnos de la universidad, en la figura 1 SIU Guaraní Total, que interactúa con el Sistema Integrador, en la figura 1 Personas UNLP, para proveer información de egresados, situación académica y actividad en general. Además permite obtener reportes ad-hoc en forma rápida y eficiente [15] [16]

Mantener actualizado el Data Warehouse fue otro punto de análisis y debate. El análisis y discriminación de la información que debía estar actualizada en forma diaria y cual no fue el primer paso. En un primer momento el Data Warehouse se actualizaba cada semana. El costo de esta actualización era muy elevado en cuanto a tiempo de procesamiento y calidad de la información obtenida. Es así como se comienza a analizar herramientas para optimizar el proceso de ETL. Pentaho BI, herramienta de software libre para BI es la seleccionada por el MEN para su grupo de trabajo y es adoptada por el equipo de Sistemas Académicos. Es así como se utiliza Kettle para contar con una vista centralizada de estudiantes, carreras y planes de estudio y egresados actualizada en forma diaria. El análisis de las variables a mantener en el Data Warehouse y el grado de actualización de cada una de ellas, evidenció la necesidad de modificar los sistemas de gestión para poder trasladar al DW sólo los últimos cambios realizados, día a día. Uno de los problemas detectados en la puesta en producción de esta etapa fue el alto índice de correcciones de datos residentes en los sistemas de gestión. Esto es producto de la visualización de la información por parte de los estudiantes a través de la Web y la integración con otros sistemas que requieren de datos confiables y de calidad. Esta realidad llevó al equipo a planificar actividades específicas relacionadas con calidad de datos.

Por otro lado, la aplicación de técnicas de minería de datos permitirá identificar los diferentes perfiles de los estudiantes y egresados, ayudando a comprender mejor su comportamiento en las plataformas virtuales y las redes sociales, para implementar distintas propuestas educativas, por ejemplo aquellas que los asistan para completar sus estudios como es el caso de la Facultad de Informática [17]. El estudio sistemático de los estudiantes desde diferentes perspectivas se puede encontrar en numerosos

trabajos llevados a cabo en el país y en el mundo [18] y [19] Sin embargo es interesante complementar estos análisis incorporando idiosincrasias propias de la universidad local y de una Facultad en particular. Los primeros avances en esta línea se están llevando a cabo utilizando Weka *W*aikato *E*nvironment for *K*nowledge *A*nalysis[20], es una herramienta visual de libre distribución (licencia GNU) desarrollada por un equipo de investigadores de la universidad de Waikato (Nueva Zelanda). La información utilizada en el proyecto de minería que se está llevando a cabo actualmente toma de los datos almacenados en la plataforma virtual y el sistema de bibliotecas, y distintas vistas del DW y utiliza Weka para aplicar las técnicas de minería que permitan obtener conocimiento de estos mares de información. El grado de aprovechamiento de la información minada por parte del usuario final depende en gran medida de una correcta visualización y una interfaz amigable de interacción. Por este motivo se prevé el desarrollo de aplicaciones ad hoc para facilitar las consultas de los usuarios finales.

Asimismo, la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Esta tecnología se encuentra en pleno desarrollo, encontrando soluciones open source que es necesario estudiar e investigar en forma sistemática para obtener resultados comparativos que sean de utilidad. Podemos mencionar aquí herramientas que realizan analytics en memoria, por ejemplo el streaming processing que realizan empresas como Linkedin, Groupon a través de aplicaciones como Storm[21] y Kafk[22]; o Drill[23] y Dramel[24] para la exploración de datos. D3[25] es otra aplicación muy poderosa para crear tableros interactivos en forma rápida y eficaz visualmente.

El volumen de información gestionado actualmente en los distintos sistemas de gestión académica, sumado al estudio del comportamiento de los alumnos en las redes sociales y las plataformas virtuales, hacen imprescindible encaminar investigaciones en esta dirección. Además permitirán extender el análisis a datos de toda la universidad y facilitar la integración las bases de datos abiertas de organismos públicos aplicando las técnicas y herramientas analizadas sobre esta temática.

3 Integración de Sistemas

En TICAL 2012[26] se hizo referencia a las soluciones desarrolladas ad-hoc para la gestión operativa de la Universidad. Por ejemplo el CMS Choique de código abierto con licencia GNU liberado en el año 2012, el sistema de gestión de Licencias Médicas, de Expedientes y de Becas entre otros , cada uno de ellos con su usuario y clave de registro que se validan contra un determinado dominio. Los empleados de la UNLP pueden tener distintos roles, y para cada rol acceder a distintos sistemas. Por ejemplo un docente puede acceder a la plataforma virtual que utiliza su facultad así como también al sistema de gestión de alumnos y consultar su recibo de sueldo, solo

por citar algunas aplicaciones. A su vez esté docente puede desempeñarse en alguna oficina de la unidad académica y acceder a un sistema para realizar su trabajo, también con usuario y clave. Es así como los roles de una persona se diversifican así como también la cantidad de usuarios y claves que debe recordar. Es así como se hizo imprescindible implementar un mecanismo de control de acceso centralizado o Single Sign On, a un conjunto de aplicaciones relacionadas pero independientes, que dialogan entre ellas a través de Internet, más allá de las tecnologías subyacentes propietarias o no. Un usuario se registra en una de las aplicaciones y gana acceso a todas las demás aplicaciones relacionadas. En el ultimo año se trabajó en migrar el registro de estas aplicaciones para soportar la autenticación centralizada basada en SSO. Como mecanismo se adoptó SAML Security Assertion Markup Language, protocolo estándar para la comunicación de identidades a través de Internet. Es un mecanismo de Single Sign On basado en XML, para comunicar autenticación o identidad, derechos o permisos y atributos de un usuario entre distintas entidades. [27] La versión 2 de este protocolo, liberada en el año 2005, es un conjunción entre distintas iniciativas de OASIS Organization for the Advancement of Structured Information Standards y Liberty Alliance Federation Framework. SimpleSAML[28] es una implementación del protocolo, que se ocupa de la autenticación, escrita en PHP Como proveedor de identidad se implementa el Sistema Integrador de Datos de la UNLP, en la figura 1 Personas UNLP. Este sistema permite centralizar la información de alumnos, docentes, no docentes y autoridades mediante la integración de datos que provienen de los siguientes sistemas: SIU-Guaraní, Sistema Integrado de Registro de Alumnos y Sistema de Personal (liquidación de sueldos). Esta base de datos es utilizada para alimentar a otros sistemas desarrollados por el CeSPI que tienden a facilitar la administración de los recursos humanos de la Universidad Nacional de La Plata.

El Sistema Integrador de Datos busca reunir toda la información de la comunidad universitaria para normalizarla, articularla y hacerla accesible. El sistema responde a los problemas de administración de recursos humanos que están teniendo todas las organizaciones. Con este desarrollo se intenta consolidar la información para evitar errores y repeticiones. Actualmente la información es utilizada por otros sistemas desarrollados en el CeSPI :

- Sistemas de salud universitaria: que se compone de dos aplicativos destinados al área de salud y a los profesionales médicos de la UNLP.
- Sistema de Libretas Universitarias: se encarga de la digitalización de las libretas sanitarias estudiantiles y el seguimiento sanitario e historias clínicas de los alumnos.
- Sistema de Carpetas Médicas y Controles Periódicos: facilita la gestión de solicitudes de carpetas médicas del personal docente y no docente de la UNLP. Las carpetas médicas se solicitan a través de Internet posibilitando que esta actividad tome un carácter centralizado. El médico recibe una planilla con todos los datos del paciente y un plano, realizado a través de Google Maps, para que ubique fácilmente el domicilio pertinente. Finalmente, este sistema permite llevar un registro de los controles periódicos.
- Sistema de inscripción a Becas: Este sistema, desarrollado para la convocatoria a becas de noviembre de 2010, fue el primero en integrarse al

Sistema Integrador de Datos. El mismo facilita la asignación de las becas que otorga la Dirección de Asuntos Estudiantiles de la UNLP a los estudiantes que las requieran. Incluye al sistema de Albergue estudiantil gestiona el ingreso y egreso de los estudiantes al Albergue Estudiantil de la Universidad Nacional de La Plata.

- Sistema de recibo de sueldos

Actualmente, se están estudiando distintas estrategias y herramientas para garantizar la calidad y disponibilidad del Sistema Integrador de Datos, por ejemplo técnicas de Big Data y herramientas de ETL alternativas a Kettle de Pentaho BI

Además, se amplían los servicios ofrecidos a través de la API Rest que consulta los sistemas de gestión académica SIU Guaraní. La primera experiencia de integración con el sistema de gestión de Títulos ha abierto la puerta a la integración con otros sistemas como las Libretas Sanitarias y la inscripción a cursos de postgrado de la Especialización en Docencia Universitaria.

Como mencionamos anteriormente, las redes sociales constituyen una realidad que no es ajena a la comunidad universitaria. En la Facultad de Informática se ha desarrollado el módulo TAM, Twitter Activity Module https://github.com/mcharnelli/moodle-module_twitter que permite relacionar un curso de Moodle con una cuenta de Twitter y a su vez con una cuenta de Facebook, utilizando el protocolo OAuth[29] Asimismo, el CMS Choique permite incluir bloques de Twitter como parte de los portales que administra.

Es interesante entonces estudiar estos bordes y la información que fluye hacia y desde las redes sociales utilizando herramientas como HootSuite[31] y TweetDeck[32] a fin de obtener y generar información significativa para la gestión.

4 Conclusiones

Las ventajas de la interoperabilidad y la potencia/versatilidad de las herramientas de extracción para construir repositorios intermedios son dos de los ejes que permiten expandir los horizontes de los sistemas de la UNLP y crear nuevas funcionalidades como describe el artículo.

En particular se favorece la interacción entre los distintos sistemas de la propia universidad, así como interactuar con los sistemas del SIU que consolidan información a nivel Ministerial y respetar los estándares de interoperatividad de las principales redes sociales como Google+, Facebook y Twitter.

La generación de repositorios intermedios permite proveer nuevos servicios gerenciales y de servicio de consulta sin afectar la performance de aplicaciones masivas como el sistema de alumnos.

La interconexión de los sistemas por otra parte simplifica procesos administrativos, ahorra papel y genera una auditoria automática en procesos críticos de la Universidad como emisión de diplomas

La evolución de la tecnología ofrece potencialidades que cuando se incorporan a los sistemas en uso (legacy) los agiornan y permiten no solo incorporar nuevas funcionalidades sino también simplfiicar accesos, cantidad de pasos e interfaces.

Referencias

[1] Portal de la Universidad Nacional de La Plata. http://www.unlp.edu.ar/institucional Última visita 28 de abril 2013
[2] http://cespi.unlp.edu.ar Última visita 27 de abril 2013
[3] http://www.tuv.com/es/argentina/home.jsp Última visita 20 de abril 2013
[4] http://unlp.edu.ar/articulo/2011/11/17/anuario_de_indicadores_2012 Última visita 20 de abril 2013
[5] http://siu.edu.ar Última visita 22 de abril 2013
[6] http://www.siu.edu.ar/nuestras-soluciones/gestion-academica-2/siu-guarani-2 Última visita 27 de abril 2013
[7]http:// moodle.org Última visita 25 de abril 2013
[8] http://www.cespi.unlp.edu.ar/meran Última visita 27 de abril 2013
[9] www.koha.org Última visita 26 de abril 2013
[10] http://www.pentaho.com/ Última visita 28 de abril 2013
[11] www.spagobi.org/ Última visita 28 de abril 2013
[12] http://www.talentanalytics.com Última visita 27 de abril 2013
[13] http://www.ideasoft.biz/ Última visita 26 de abril 2013
[14] The Data Warehouse Lifecycle Toolkit, 2nd Edition: Practical Techniques for Building Data Warehouse and Business Intelligence Systems. John Wiley & Sons, 2008
[15]http://www.ing.unlp.edu.ar/institucional/difusion/2012/ingreso_retencion Última visita 10 de marzo 2013
[16] http://www.econo.unlp.edu.ar/caracterizacion_aspirantes Última visita 10 de marzo 2013
[17] http://www.info.unlp.edu.ar/articulo/2010/7/19/info_secretaria_academica Última visita 12 de marzo 2013
[18] http://www.datapr/ix.com/blogs/gpautsch/resumen-mi-tesis-miner-datos-aplicada-lisisdeserci-n-carrera-analista-sistemas-compu Última visita 12 de abril 2013
[19] www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf Última visita 19 de marzo 2013
[20] http://www.cs.waikato.ac.nz/ml/weka/ Última visita 29 de abril de 2013
[21] storm-project.net Última visita 27 de abril de 2013
[22] http://kafka.apache.org/design.html Última visita 27 de abril de 2013
http://www.ibm.com/developerworks/ssa/local/im/queesbig-data/index.html Última visita 25 de abril de 2013
[22] http://kafka.apache.org/design.html Última visita 25 de abril de 2013
[23] http://gigaom.com/cloud/for-fast-interactivehadoopqueries-drill-may-be-the-answer/ Última visita 25 de abril de 2013
[24]http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/36632.pdf Última visita 25 de abril de 2013
[25] http://d3js.org Última visita 25 de abril de 2013
[26] http://tical_2012.redclara.net/es/presentaciones.html Ultima visita 25 de abril de 2013
[27] http://en.wikipedia.org/wiki/Security_Assertion_Markup_Language Última visita 25 de abril de 2013
[28] http://php.net/manual/es/book.simplexml.php Última visita 25 de abril de 2013
[29] https://github.com/mcharnelli/moodle-module_twitter Última visita 25 de abril de 2013
[30] http://oauth.net/ Última visita 20 de abril de 2013
[31] http://hootsuite.com/ Última visita 25 de abril de 2013
[32] http://tweetdeck.com/ Última visita 26 de abril de 2013

EXPIRIENCES WITH EDUCATIONAL ROBOTIC

Anibal Lopes Guedes[1], Fernanda Lopes Guedes[2]

[1] Professor at Univerisidade Federal da Fronteira Sul – UFFS.
Master in Computer Science.
[1] Professor at Instituto Federal Sul-Rio-Grandense – ISUL.
Master in Computer Science.

{anibalguedes, fernandalguedes}@gmail.com

Abstract. Educational Robotics has come to prominence in recent years because it allows to articulate a more playful and interactive teaching. It works the abstract in concrete basis, and this way stands as a new methodology of teaching and learning. Therefore, this article seeks to present educational initiatives that use robotics in elementary schools, located in western Santa Catarina and northern Rio Grande do Sul. For that, we used the Lego Mindstorms NXT robotic kit. As result, it was found that the technology enables the insertion, integration, interaction, discussion and cooperation between students, teachers and employees, which somehow permeates individual and collective development, providing opportunities for improvements in the educational processes.

Keywords: Educational Robotics. Constructionism. Robotic. Automation. Lego Mindstorms NXT.

1 INTRODUCTION

Education leads us to a series of situations, practices and policies that bind the area itself [2]. In this perspective, it is possible to highlight the importance of transformational technologies that offer educational processes, serving as support for the proposals developed, contributing to changes in the social and cultural dynamics of individuals.

In this sense, [3, 4, 5] is stated that it is up to the task of the educators to plan and introduce such technologies in school life.

The computer is not only to perform human tasks such as to add, process and teach, but also to require the development of cognitive and meta-cognitive skills of each individual through learning situations that enable a better understanding of the world in which we live in[6].

The author states that technologies have a transformative approach, since: They alter the structure of interests of each individual; Change the thinking of each individual; Alter the nature of the community.

Corroborating this proposal, [7] he states that technology should promote the development of basic skills and cognitive abilities of their users, explore learning in an interactive and playful way, allowing people to new educational processes, new experiences, new discoveries and new ways of learning. Thus, the robot is attractive as a means [8], "invites teachers and students to teach / learn / discover / invent in collective processes, capable of connecting abstraction and concrete world."

Through them, you can explore the area of robotics in an educational manner, coming to join efforts to make school life more challenging, creative and focused on the processes of teaching and learning. The use of robotics in the classroom, according to [9], provides that "teachers escape the blackboard and that lessons become more dynamic thereby arousing the curiosity of students," setting up what could be called technological literacy.

In Brazil, projects conducted by the Educational Robotics, are yet isolated initiatives. There is still a look that directs efforts to robots that can support the school setting as a means to include the Computer within other disciplines such as Mathematics, Physics, Biology, Portuguese and others [5]. From their research on the use of educational robotics, [8] it pointed out that: "Countries such as the Netherlands and Germany have already robotics [...] 100% of public schools. England, Italy, Spain, Canada and the United States go in the same direction. Some Latin American countries have adopted their first strategies nationwide. This is the case, for example of Mexico and Peru."

[10] claim that the university is a place of production significant social and technological initiatives that are carried out in this case study on Robotics in Education, in order to: Knowing closely the social reality of the public attended in order to modify it; Provide the qualification of the citizen; democratize access to knowledge gained to improve the quality of life of citizens; Encourage scientific research; Promoting citizenship and democratic values to the different social actors who are involved directly and indirectly in the shares.

Thus Robotics Education opens unexplored possibilities for the field of education and to the field of research, transforming educational settings.

So, this paper aims to present considerations of Robotics, Robotics Education, present the Lego Mindstorms NXT robotic kit, and show initiatives in education level and extent applied with students of the initial series.

Finally, the text helps to diversify the work done by the teacher, and this will have available a means versatile, able to cause a modification of traditional culture and organization of the school, contributing the learning of their students.

2 ROBOTICS

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

Robotics is a branch of technology that includes mechanics, electricity, electronics and computing, which deals with systems composed of mechanical parts and machines: automatic and controlled by integrated circuits, making mechanical motorized, controlled manually or automatically by electrical circuits [11].

The term robotics was created by science fiction writer Isaac Asimov in his novel "I, Robot", 1948 [11]. The word “robot” was first used in 1921, a play that was titled RUR - Russum's Universal Robots, Czechoslovakia, written by Karel Capek. In Czech, the word “robota” means work and was used in the sense of a robot machine to replace human labor.

The difference between a computer and a robot is districted by its power to interact with the world. The computer does not start without human operation. Thus, robot is considered an intelligent mechanism that works autonomously [11].

From the 80, the Robotics is advancing at great speed and, among numerous projects, ASIMO, initiated in 1986 by the Honda Motor Company, to get highlighted. Contrary to what it may seem, his name was not created in homage to the science fiction writer Isaac Asimov, but is derived from Advanced Step in Innovative Mobility. Just like ASIMO, the Qrio, Sony, and Robonaut robot created by NASA to assist astronauts on the International Space Station performing extra vehicular activities, are also quite relevant. The three are cited as humanoid robots designed to interact with humans [12, 13].

Besides these, [14] presents the home robot NAO, which can recognize the face of a person and interact in conversation. RI-MAN and RIBA II are other examples of robots that provide assistance to a person’s body.

Something that caught our attention is that “the Internet has proved valuable in providing access to similar lines of research, sharing open source materials and facilitated exchange of opinions and resources, which benefits the improvement of technologies.” [14, 15].

Thus, the next section presents considerations regarding one of the areas that benefit most from the Robotics, in this case, Education.

3 EDUCATIONAL ROBOTIC

The technologies are important tools for studying and research in the learning process, as they provide conditions to both teachers and students working from themes, projects and extracurricular activities. [16] states that the computer is a medium that develops attention, perception and creativity.

Corroborating this idea [17] states that the computer is like “a machine [...] that allows to test ideas or hypotheses, leading to the creation of a world abstract and symbolic that at the same time allows you to enter different forms of operation and interaction between people.”

This is a device that is increasingly diverse in functions, contributing significantly to an increase in productivity, cost reduction and optimization of product quality and services.

For this reason that the school should support projects where the computer presents real situations to students in order to make your learning fun and engaging, the example cited is Educational Robotics.

The Robotics Education and teaching, named, “[...] encourages creativity students due to its dynamic nature, interactive and even playful besides serving as a motivator to stimulate students' interest in traditional teaching.” [18].

It is characterized as an environment in which students can "program" and "build" your robot. "The ease of installation and programming of robots, articulated piece sets and intuitive programming interfaces can be identified as factors that [...] put in a field of robotics accessible to educational purposes." [19].

The advantages of Educational Robotics are very significant. Among the benefits are: interdisciplinary, the expansion of content already worked in the classroom and, what is more important, the learning achieved through group work, since the study phase. Principles of teamwork and cooperation, which are required in professional practice, skills are developed in students from the Robotics projects [20].

In this way, we use the cognitive model of the Theory of Multiple Intelligences (MI), proposed by Gardner [21] which describes for the coexistence of multiple intelligences. Thus, relating the theory of IM with Educational Robotics have:

- linguistic intelligence - students can express themselves through words as a robotic experiment passed or that was developed;
- logical-mathematical intelligence - students can reason about how to solve the problem by means of the robot and how to program it;
- spatial intelligence - the student can understand how the pieces fit the robot to assemble the robotic structure (visual perception);
- kinesthetic intelligence - the student can, by means of somatic sensations obtained through sensors, articulate his/her robot to obtain elements for analysis;
- musical intelligence - the student can, through rhythms and melodies, listening posts, sounds and music via robot;
- inter-personal intelligence - work as a team with the involvement of the group, assembling, programming and the use of the robot;
- intra-personal intelligence - the ability to act adaptively meets the challenge presented;
- naturalist intelligence - the ability of interaction between the student and the environment, in this case, using recyclable artifacts.

Projects addressing the theme of Educational Robotics develop in several segments, geared mostly to high school and vocational education. [21] states that there are few projects articulated with the elementary school. The author states that there are few institutions in fundamental level that include content related to technology education into their curricula. Among the projects using robotics as learning environment IN Brazil:

- Technological Education - it's a project that allows experience knowledge in areas such as Physics, Biology, Mathematics and Language, through the assembly and programming of robotic kits. The project is carried out in primary and secondary schools in Feira de Santana, Bahia [23];
- Educational Robotics in High School - its experiments are performed between the disciplines of Geography, Mathematics and Programming by robots. Project developed in Blumenau - Santa Catarina [24];
- Educational Robotics in UCA - the project aims to analyze "technical activity as symbolic and creativity from the field of possibilities of modeling and programming in the context of design prototypes UCA." [25]. The project took place in Porto Alegre (Rio Grande do Sul) with public and private schools from elementary and high school.

At the level of the technological resources available in the market to work with robots, there are free solutions and cost as the Arduino, the GoGoBoard the xBot, among others. Already in level sets (kits) and manufactured commercial, there is the Lego Mindstorms NXT, the Fischertechnik, Parallax, among others.

[20] points out that the basic difference between commercial complexes in relation to free ones, are: number of parts (gears, motors, etc.). In this aspect of low cost kits contain a limitation on the number of its parts; Collection of alternative materials. In this aspect both kits cost as commercial or manufacturing allow the customization of parts; Technical knowledge of electronic and mechanical modelling. In this aspect kits that require low cost to learn a specific programming in accordance with the kit purchased.

Based on the questions listed by [20], in the next section, we present some considerations about the commercial kit Lego Mindstorms NXT. The kit was chosen because it contains a simple programming interface and iconic, many examples available on the website http://www.nxtprograms.com plus numerous parts such as gears, motors and sensors support [26].

4. PROPOSALS DEVELOPED WITH CHILDHOOD EDUCATION AND ELEMENTARY SCHOOL

As the authors [28], the school is one of the most important social institutions, because it is the link that mediates the interaction between the individual and society, allowing that the child can take ownership of values and social models, with direct repercussions on their autonomy.

Therefore, the technology is part of this link, as it allows them to adopt actions that facilitate the educational process. In this sense, the Educational Robotics in schools aims to provide students with the awakening of logical reasoning, creativity, autonomy in learning, understanding concepts and improve coexistence group, treating cooperation, planning activities and tasks [29].

Thus, in Section III set out examples in Brazil's level of technological use of robotics in the classroom. Obviously, corroborating [8] the area is still shy and lacking in trained professionals at both technological as teaching. However, it is at this time showing initiatives experienced by researchers in level of education and outreach. The proposals are summarized here from the work: [30, 31, 32].

In both projects we used as instruments to direct observation of school and classroom intervention, a second time.

The observation has the advantage of identifying the facts directly, without any intermediary, as indicated by [33], it is a simple methodology and systematic. Thus, the first step involved the analysis of impactual technologies, especially robotics in schools.

Therefore, there was an informal contact with representatives of schools, these included teachers and principals in order to realize their positions through technologies. This contact allows us to understand how the robot can interfere with the educational process, since, according to statements summarized, “[...] it is a mechanism to aid the teacher, it enhances the process of child development, is

attractive, is a motivating tool that arouses interest, changes the dynamics of teaching and learning, motor elaborates on the student, a different way to share their knowledge with students and can open many doors for these children who are starting school life."

The next step in the study provided the Lego Mindstorms NXT robotic kit in order to analyse and understand their use, operation and programming. The robot was chosen because it is a line of Lego toys released by the company, focused on technology education, and for being a technology widely used in the process of teaching and learning in schools.

[27] posits that humans learn best when they are engaged in building something that he can show to other people and that is meaningful to him. These computing environments, especially Robotics, contribute to this way of thinking constructionist, because students engage and interact with the development of projects.

The study on the Mindstorms Lego NXT kit is considered with step observation of classes in schools - presented in the sequence, where the idea is: a) to investigate the teaching practice, directly observing how the activities are developed, explored and its relevance to students; b) analyse the relationship between student and teacher in the classroom.

From the analysis of reality and the school plan for each school, went to trial activities in technological level. For this, we used the flowchart first proposed by [30]. In this case, the flowchart is composed: Name - name of the experiment; Objective(s) - highlight the educational purposes; Discipline(s) - experiment (s) intended; Materials used - detailed description of physical materials for the experiment; Description of the stages - description of the step by step to the realization of the activity.

From the structure of the experimental projects using the Lego Mindstorms NXT robotic kit, went up to the stage of intervention in school. The intervention process has the intention to facilitate the "problematic at collective practices of training and enhancing the production of a new think / do education." [34].

The intervention included the formation of teachers and other legal representatives of schools, through workshops, so that they can meet the robotic kit and check usability. After that, it went to the training and validation of the project with the students.

Finally, the initiatives have great concern for the educational processes of human development. Following considerations and presenting peculiarities of each initiative indicated with the results obtained.

4.1 Iniciative Kerber

The initiative [30] was applied in Ambial Project, Escola de de Educação Básica São Lourenço, in the municipality of Iporã do Oeste-SC. The project aims to promote the inclusion of pedagogical actions socio-environmental, mitigating the problem of hunger, education and sustainability of the students.

[30] applied his work against the shift of student activities. As the study areas, highlight Sciences, Mathematics and History. Among the experiments developed and applied:

• Interactive Game - the idea of the game is to interact with the Lego Mindstorms NXT robot. In the game there is a specific symbol that appears in the central controller. Thus, the student is invited to interact selecting between the buttons contained in the controller in order to ascertain where the symbol is showing up randomly;
• Commanding the vehicle by remote control - vehicle was developed, similar to the transport of the pieces of Lego Mindstorms, which moves around on two wheels, which are located in front of the application and the lower end of which was affixed a wheel lowest, responsible for the direction of the vehicle;
• Distance in centimeters from one point to another in a straight line - it's an application which calculates the distance in centimeters from one point to another in a given line segment, displaying them in real time on the display of the central controller;
• Hieroglyphics - with parts and programming Lego together in the shape of a vehicle, it is the representation of some symbols (hieroglyphics), so that students can understand this writing in a fun and interactive way, using the History and Mathematics to encourage logical reasoning;
• Calculation of area and volume - application which calculates the area and volume of objects that the students will choose, displaying them in real time on the display controller;
• Representation of the solar system - development of a framework for a rotary level, with the parts and programming Lego set, which makes movements similar to the solar system. The proposed design allows three variations on an experimental level. The first would be the Earth revolving around the Sun, and the second would be the planets rotating around the Sun, the Moon would be the third turning on the Earth.

4.2 Initiative Tosini and Holz

The initiative [31] was applied at the Escola Estadual Catharina Seger, located in the municipality of Palma Sola-SC, with students in the class multi-seriate the first and second years.

The initiative [31] considered the use of Bluetooth technology in parallel to the use of Lego Mindstorms NXT robotic kit.

The Bluetooth allows communication via radio signals from high frequency between computers, smartphones, cell phones, mice, keyboards, headsets, printers and other devices.

To make this possible, it took two robotic kits. The first, called the master is the device that creates the connection, while the other, called the slave performs the action.

The project was carried out taking into account the areas of:
• Mathematics - with understanding the geometric shapes, basic arithmetic operations of addition and subtraction, colors;
• Sciences - healthy eating habits through educational activities that inform and motivate individual choices.

At the trial there was a special student with a mild mental retardation. The child that contained aggressive behavior with the teacher and classmates, the experiments performed satisfactorily and felt motivated to help their classmates. At work there was not a closer study on the use of technology in special education.

4.3 Initiative Zarpelon, Tortelli and Bieniek

The initiative extension was applied to three schools, two located in the municipality of Erechim-RS (Escola Estadual de Ensino Médio Irany Jaime Farina and Escola Municipal de Educação Infantil Dom João Aloisio Hoffmann) and the other in Passo Fundo-RS (Escola Municipal de Ensino Fundamental Georgina Rosado). Participated in the project students from kindergarten and first grade of elementary school.

In the phase of experimentation, was developed a board game. [15] believes that a play can develop play behavior, anticipating the behavior of the child, adult and old. "[...] The game is working well, the duty, the ideal life." That is, pervades the individual's independence.

In the case of the project, the game involves environmental issues, since it is one of the crosscutting themes of education and research focused in schools. The goal is to develop logical thinking through differentiation of geometric figures, as well as its dimensions, color differentiation, the interaction with the world technological and environmental awareness, all these issues will appear in the course of the board. It should be noted that the board game involving environmental issues is just one of the indications to be used in the process.

From the results obtained it can be seen that the enormous interest on the part of students and teachers, who also supported the project, the integration of robotics in the school learning environment. Students showed greater attention, concentration and pledged to develop the proposed activities on the board.

5. CONCLUSIONS

From the results obtained with the Lego Mindstorms technology, the researchers realized the advantage of the kit in the learning of both students and teachers; it provides creating imaginative concrete structures, ranging from humanoid replicas of animals, vehicles, among others.

In contrast to this, it is clear also that the robotic kit is still expensive in Brazil and that it can be a complicating factor, since public schools depend on state and local budgets.

In order to meet the social reality of the public attended, schools attended were visited, to see how the students are in the classroom, what activities are developed by teachers, the main difficulties faced by students in their learning process, if there are initiatives in schools with the use of technologies. It was found that, lacking training processes that encourage the use of computers in school, particularly robots, with "small" (personal lines of teachers consulted).

[30] highlights the need for an overhaul of the school curriculum, teacher training and school representatives so that they can work properly interdisciplinarity that technology can provide.

The initiative also influenced the relationship between the groups, allowing greater communication between students and teachers, which in a way, was distant.

Now [31, 32] highlight the involvement of staff (students and teachers) in the process of experimentation. They also indicate that the curriculum reform and the training of teachers, raised by [30], as necessary, relevant and urgent in school.

When asked about how they imagined a robot for some robot were only those with humanoid forms, or had other idea would be a robot. And when asked about the proposed activities, the students would like other activities that were proposed [30, 31, 32].

Therefore, the university has the main role to change the reality pointed out by researchers, either with incorporating technological, scientific, educational process, professional skills in order to create a more just society that promotes the development of individuals who make it part.

REFERENCES

01. BRANDÃO. Carlos R. O que é educação. 33. ed. São Paulo, SP: Brasiliense, 1995.
02. CHARLOT, Bernard. A pesquisa educacional entre conhecimentos, políticas e práticas: especificidades e desafios de uma área. Revista Brasileira de educação. v. 11, n. 31, p. 7-18. jan./abr. 2006.
03. OLIVEIRA, Ramon de. Informática Educativa. Campinas, SP: Papirus, 1997. 176 p.
04. PEIXOTO, Joana. Metáforas e imagens dos formadores de professores na área da informática aplicada à educação. 2007. Educ. Socio., v. 28, n.101, p. 1479-1500. Disponível em: <http://dx.doi.org/10.1590/S0101-73302007000400011>. Acesso em: 22 abr. 2012.
05. CRUZ, M. E. J. K.; et. al. Formação prática do licenciado em Computação para trabalho com Robótica Educativa. In: XVIII Workshop em Informática na Educação (SBIE), São Paulo, SP, 2007.
06. SANCHO, Juana Maria. De Tecnologias da Informação e Comunicação a Recursos Educativos. In: SANCHO, Juana Maria. et. al. Tecnologias para transformar a Educação. Porto Alegre, RS: Artmed, 2006. p. 15-40.
07. CORREIA, Secundino. Inteligência Emocional e Robótica na Educação. Revista Perspectiva, 2008. Disponível em: <http://bica.imagina.pt/2008/inteligencia-emocional-e-robotica-na-educacao/>. Acesso em: 07 abr. 2013.
08. QUINTANILHA, Leandro. Irresistível robô. Revista ARede, ed. 34, mar. 2008. Disponível em: <http://www.arede.inf.br/edicao-n-34-marco-2008/3920-irresistivel-robo>. Acesso em: 07 abr. 2013.
09. PRADO, José Pacheco de Almeida. Robôs estarão disponíveis para estudantes brasileiros. 2008. Disponível em: <http://www.acessasp.sp.gov.br/2008/02/robos-estarao-disponiveis-para-estudantes-brasileiros/>. Acesso em: 11 ago. 2012.
10. TREVISOL, J. V.; CORDEIRO, M. H.; HASS, M. (Org.). Construindo agendas e definindo rumos. Chapecó, SC: UFFS, 2011.
11. MURPHY, R. R. Introduction to a robotics. Cambridge: The Mit Press, 2000.
12. AYRES, Marcelo. Conheça a história dos robôs. Disponível em: <http://tecnologia.uol.com.br/ultnot/2007/10/01/ult4213u150.jhtm>. Acesso em: 11 abr. 2013.
13. ROBOLIVRE. História da Robótica. Disponível em: <http://robolivre.org/conteudo/historia-da-robotica>. Acesso em: 11 abr. 2013.

14. BAKER, James. Robótica de Última Geração. Como Funciona, São Paulo, n. 10, a. 1, p.44-47, 2013.
15. CHATEAU, Jean. O jogo e a criança. São Paulo: Summus, 1987.
16. LIANO, José Gregorio de; ADRIÁN, Mariella. Formação Pedagógica: A informática Educativa na escola. São Paulo, SP: Loyola, 2006.
17. GONÇALVES, Maria de Jesus. Linguagem e tecnologia. In: DELIBERATO, Debóra. Comunicação alternativa: teoria e prática. São Paulo, SP: Memnon Edições Científicas, 2009.
18. ROCHA, Sinara Socorro Duarte. O uso do Computador na Educação: a Informática Educativa. Revista Espaço Acadêmico, n. 85, jun. 2008. Disponível em: <http://www.espacoacademico.com.br/085/85rocha.htm>. Acesso em: 06 abr. 2013.
19. GOMES, Marcelo Carboni. Reciclagem Cibernética e Inclusão Digital: Uma Experiência em Informática na Educação. In: LAGO, Clênio (Org.). Reescrevendo a Educação. Chapecó, SC: Sinproeste, 2007. 202 p.
20. LOPES, Daniel Queiroz. Brincando com robôs: desenhando problemas e inventando porquês. Santa Cruz do Sul, RS: EDIUNISC, 2010.
21. GROCHOCKI, Luiz Rodrigo; SILVA, Rodrigo Barbosa e. Robótica Educacional. Guarapuava, PR: Roboticaeducacional.com.br, 2009.
22. ARMSTRONG, Thomas. Inteligências Múltiplas na sala de aula. Porto Alegre, RS: Artmed, 2001.
23. GUIMARÃES, Gleidson Carneiro. Robótica: Espaço interdisciplinar de estímulo às inteligências múltiplas. Revista do Professor, a. 24, n. 96, out./dez. 2008. Porto Alegre, RS.
24. BENITTI, Fabiane Barreto Vavassori .; et. al Experimentação com Robótica Educativa no Ensino Médio: ambiente, atividades e resultados. In: XV Workshop sobre Informática na Escola (WIE), Bento Gonçalves, RSl, 2009.
25. LOPES, Daniel Queiroz; FAGUNDES, Léa da Cruz; BIAZUS, Maria Cristina V .Robótica Educacional: técnica e criatividade no contexto do Projeto Um Computador por Aluno. In: XIX Simpósio Brasileiro de Informática na Educação (SBIE 2008), Fortaleza, CE, 2008.
26. FORD JR., Jerry Lee. Lego Mindstorms NXT 2.0 for Teens. Boston, MA: Cengage Learning, 2011.
27. PAPERT, Seymour. A máquina das crianças: repensando a escola na era da informática. Porto Algre: Artes Médicas, 1994.
28. BOCK, A. M. B.; FURTADO, O.; TEIXEIRA, M. L. T. Psicologias: Uma introdução ao estudo da Psicologia. 14. ed. São Paulo: Saraiva, 2008.
29. PIO, J. L. de S.; CASTRO,T. H. C.; CASTRO JUNIOR, A. N. A Robótica Móvel como instrumento de apoio à Aprendizagem de Computação. In: XVII Simpósio Brasileiro de Informática na Educação - SBIE, Brasília-DF, 2006.
30. KERBER, Fábio Matias. Usando a Robótica como meio Educativo. Trabalho de Conclusão de Curso - Curso de Sistemas de Informação, Universidade do Oeste de Santa Catarina, 2009. 86 p.
31. TOSINI, Juliana; HOLZ, Franciane de Cassia. O emprego da tecnologia Bluetooth e robô Lego Mindstorms no Aprendizado de crianças. Trabalho de Conclusão de Curso - Curso de Sistemas de Informação, Universidade do Oeste de Santa Catarina, 2010. 63p.
32. ZARPELON, Mirian Cátia; TORTELLI, Luana; BIENIEK, Gregori Betiato. O uso da Robótica nos processos educativos de alunos da Educação Infantil e Ensino Fundamental. Projeto de Extensão, Universidade Federal da Fronteira Sul, 2013.
33. GIL, Antonio Carlos. Métodos e técnicas de pesquisa social. 4. ed. São Paulo: Atlas, 1994.
34. ROCHA, Marisa Lopes da; AGUIAR, Katia Faria de. Pesquisa-intervenção e a produção de novas análises. Psic. cienc. prof., Brasília, v. 23, n. 4, dez. 2003.

Desafíos y herramientas para la enseñanza temprana de Concurrencia y Paralelismo

Laura De Giusti[1], Fabiana Leibovich[1], Mariano Sánchez[1], Franco Chichizola[1], Marcelo Naiouf[1], Armando De Giusti[1,2]

[1] Instituto de Investigación en Informática LIDI (III-LIDI) – Facultad de Informática –UNLP
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Argentina
{ldgiusti, fleibovich, msanchez, francoch, mnaiouf, degiusti}@lidi.info.unlp.edu.ar

Abstract. Se analiza la introducción temprana de los temas básicos de concurrencia y paralelismo, de acuerdo a las tendencias curriculares impulsadas por el cambio tecnológico.
El artículo analiza la problemática a partir de la introducción de los procesadores de múltiples núcleos y las nuevas arquitecturas asociadas con Cluster, Multicluster y Clouds basados en arquitecturas multicore / GPGPU.
Se presenta una herramienta que combina un entorno visual interactivo para la programación concurrente, con el empleo de robots de demostración, especialmente para los temas de comunicación y sincronización.
Por último se analizan aplicaciones que extienden el alcance del entorno desarrollado, su aplicación en diferentes cursos y las líneas de I/D futuras en el tema.

Keywords: Concurrencia, Paralelismo, Currícula, Entorno, Multirobot, Algoritmos Concurrentes y Paralelos.

1 Introducción

La Concurrencia ha sido un tema central en el desarrollo de la Informática y los mecanismos de expresión de procesos concurrentes que cooperan y compiten por recursos ha estado en el núcleo curricular de los estudios de Informática desde la década del 70, en particular a partir de los trabajos fundacionales de Hoare, Dijkstra y Hansen [HOA78][HOA85][DIJ65][DIJ78][HAN77].
Estos conceptos se enseñaron tradicionalmente partiendo de la disponibilidad de un único procesador central, que podía explotar parcialmente la concurrencia de un dado algoritmo, en función de la arquitectura física disponible (incluso con hardware específico como los coprocesadores, los controladores de periféricos o esquemas vectoriales que replicaban las unidades de cómputo aritmético-lógico).
El paralelismo, entendido como "concurrencia real" en la que múltiples procesadores pueden operar simultáneamente sobre múltiples threads o hilos de control en el mismo instante de tiempo, resultó durante muchos años una posibilidad limitada por la tecnología de hardware disponible [HWA84][HWA93][DAS89].

En las currículas informáticas clásicas [ACM68][ACM78][ACM99] aparecían los conceptos de concurrencia en diferentes áreas (Lenguajes, Paradigmas, Sistemas Operativos) y se omitía casi totalmente el tratamiento del paralelismo, salvo al plantear los conceptos de sistemas distribuidos.
La aparición del lenguaje ADA [OLS83] a mediados de los 80 marca un hito en la evolución del tema, ya que especifica claramente en un lenguaje real los diferentes mecanismos de expresión de la concurrencia y al mismo tiempo deja clara la posibilidad de asociar los procesos ("tasks" en ADA) a diferentes procesadores físicos.
Las nuevas arquitecturas de los procesadores, que integran múltiples "cores" o núcleos en un procesador físico [GEP06][MCC08][GPG] han producido un notorio impacto en el desarrollo de la Informática, obligando a replantear el "modelo base" de un procesador. Esto ha llevado a reemplazar el formato de "máquina de Von Neuman" [GOL72] con un solo hilo de control, por un esquema como el de la Figura 1 que integra múltiples "cores" cada uno con uno o más hilos de control y varios niveles de memoria accesible en forma diferenciada [AMD09].

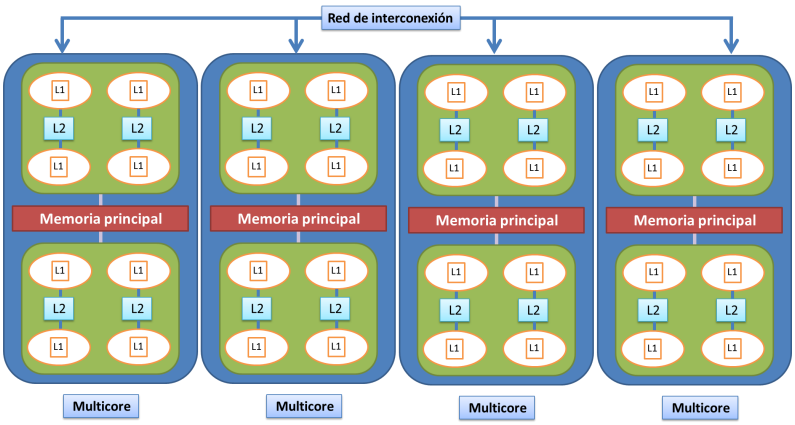


Fig. 1 Esquema de un procesador básico actual.

Al mismo tiempo, los cambios tecnológicos han producido una evolución de los temas de mayor interés en Informática, fundamentalmente por las nuevas aplicaciones que se desarrollan a partir de disponer de arquitecturas y redes de comunicación de mayor potencia y menor costo [DEG13][HOO13].
Esto ha llevado a que las recomendaciones curriculares internacionales [ACM04] [ACM08][ACM13] mencionen la necesidad de tratar los temas de concurrencia y paralelismo en etapas tempranas de la formación del alumno, dado que todas las arquitecturas y sistemas reales con los que trabajará son esencialmente paralelos.
Sin embargo, aquí aparece uno de los problemas importantes, ya que la programación paralela (y los conceptos fundamentales de concurrencia) resulta más compleja para un alumno en las etapas iniciales de su formación. Es necesario contar con nuevas herramientas que permitan abordar tempranamente el tema [CAR03][DEG12a].

2 Objetivos del entorno multirrobot en desarrollo

En este trabajo se presenta el entorno CMRE (Concurrent Multi Robot Environment) como una herramienta destinada a introducir los conceptos de concurrencia y paralelismo, con un enfoque visual e interactivo combinado con el empleo de robots físicos para la demostración de los conceptos y ejemplos de desarrollo.

2.1 Entorno visual en Concurrencia y Paralelismo

Se ha partido del trabajo desarrollado en el III-LIDI [DEG12b][DEG12c] para la enseñanza de conceptos de concurrencia en un curso CS1 buscando extender el mismo en los siguientes ejes:

- Poder declarar "procesadores" o robots virtuales que representan los "cores" de una arquitectura multiprocesador real. Estos robots virtuales pueden tener un reloj propio y diferentes tiempos para la ejecución de sus tareas específicas.
- Tener la capacidad de declarar recursos compartidos o exclusivos, incluyendo la posibilidad de tener exclusión mutua selectiva.
- Establecer objetos virtuales que representen datos básicos que se pueden contar y manipular en forma simple mediante primitivas de los robots virtuales.
- Disponer de primitivas de comunicación por mensajes sincrónicos y/o asincrónicos.
- Disponer de primitivas de sincronización por memoria compartida.

2.2 Incorporación de robots físicos al entorno visual

A los puntos anteriores se le incorpora la comunicación en tiempo real con robots físicos de la línea Lego Mindstorms EV3 [LEGa][LEGb] de modo de poder ejecutar algoritmos paralelos en el entorno, con efecto directo en los robots físicos que replican sobre el terreno el comportamiento definido por los algoritmos.
Este modelo de demostración facilita la comprensión de determinados problemas por parte del alumno, tales como los conceptos de *fairness, deadlock o inanición* [AND00].

3 Arquitectura y Primitivas de Concurrencia/Paralelismo en CMRE. Ejemplos

El entorno CMRE surge como una evolución del entorno Visual da Vinci cuyo objetivo principal fue resolver problemas donde se especifica el comportamiento de un *único robot*, el cual puede moverse en una ciudad compuesta por 100 avenidas (verticales) y 100 calles (horizontales) y es capaz de distinguir objetos (flores y papeles) y realizar operaciones con los mismos (juntarlos y/o depositarlos). Asimismo el robot puede "contar" e "informar" resultados. En la Tabla 1 se sintetiza la metáfora buscada con el nuevo entorno.

Tabla 1. Analogía entre el entorno CMRE y los conceptos de Concurrencia y Paralelismo

Conceptos de Concurrencia y Paralelismo	Entorno CMRE
Múltiples procesadores / cores	Múltiples robots (implementado con un proceso por robot)
Memoria compartida	Áreas compartidas de la ciudad
Memoria distribuida	Áreas exclusivas por robot
Memoria compartida y distribuida	Áreas parcialmente compartidas
Comunicación entre procesos por mensajes	Envío y recepción de mensajes entre robots.
Exclusión mutua sobre recursos compartidos	Bloqueo de esquinas de la ciudad.
Exclusión mutua selectiva	Acceso a áreas parcialmente compartidas.
Modelo de ejecución sincrónico	Reloj virtual sincrónico.
Arquitecturas heterogéneas	Asignar tiempos específicos a las operaciones de cada robot.
Datos locales o globales	Objetos numerables en la ciudad (flores/papeles).

En la Figura 2, a modo ilustrativo se define la estructura general de un programa en el entorno CMRE, en función del cual se especificarán sus primitivas.

```
programa areas1
areas      {defino la estructura de la ciudad}
  nombreArea1: tipoArea(Coordenada0, Coordenada1, Coordenada2, Coordenada3)
  nombreArea2: tipoArea(Coordenada0, Coordenada1, Coordenada2, Coordenada3)
robots       {defino el comportamiento de cada tipo de robot}
  robot tipo1
  comenzar
    {cuerpo}
  fin
  robot tipo2
  comenzar
    {cuerpo}
  fin
variables      {creo los robots}
   nombreVariableRobot1: tipo1
   nombreVariableRobot2: tipo2
comenzar
  {Asigno areas privadas a cada robot}
  AsignarArea(nombreVariableRobot1, nombreArea1)
  AsignarArea(nombreVariableRobot2, nombreArea2)
  iniciar(nombreVariableRobot1, PosAv, PosCa)
  iniciar(nombreVariableRobot1, PosAv, PosCa)
fin
```

Fig. 2. Estructura general de un programa en el entorno CMRE.

Tal como se mencionó anteriormente pueden resumirs las capacidades del ambiente CMRE de la siguiente manera:

- Existen múltiples procesadores (robots) que realizan tareas y que pueden cooperar y/o competir.
- El modelo de ambiente ("ciudad") en la que desarrollan sus tareas admite áreas privadas, parcialmente compartidas y totalmente compartidas. En un área privada sólo puede moverse un único robot, en un área parcialmente compartida se especifica el conjunto de robots que pueden moverse en ella y en un área totalmente compartida todos los robots definidos en el programa pueden moverse dentro de ella.
- Si se instancia un sólo robot en un área que abarque toda la ciudad, se repite el esquema del Visual Da Vinci.
- Cuando dos o más robots están en un área compartida (parcial o totalmente), compiten por el acceso a las esquinas del recorrido y a los recursos que allí existan. Para esto deben sincronizar.
- Cuando dos o más robots (en un área común o no) desean intercambiar información (datos o control) deben hacerlo por mensajes explícitos.
- La sincronización se da por un mecanismo equivalente a un semáforo binario.
- La exclusión mutua puede generarse con la declaración de las áreas alcanzadas por cada robot. Acceder a otras áreas de la ciudad, así como salir de ellas no está permitido.
- Todo el modelo de ejecución es sincrónico y permite la existencia de un reloj virtual de ciclos, que a su vez permite asignar tiempos específicos a las operaciones, simulando la existencia de una arquitectura heterogénea.
- El entorno permite ejecutar el programa de manera tradicional, o paso a paso por instrucciones, dando al usuario un control detallado sobre la ejecución del programa, de manera de poder controlar situaciones típicas de concurrencia tales como conflictos (colisiones) o deadlocks.
- En la ejecución paso a paso, el efecto de las operaciones se puede reflejar en los robots físicos, comunicados vía WI-FI.
- En el entorno, cada robot tiene asociado un estado, en el que muestra el contenido de su bolsa (cantidad de flores y papeles en el modelo), esquina donde se encuentra situado, estado actual: si se encuentra ejecutando, esperando la llegada de un mensaje, o esperando por la liberación de una esquina.

3.1 Declaración de áreas

La declaración de áreas comienza con la palabra clave **areas** y termina donde se encuentra la palabra clave **robots**. Un área de la ciudad es un subconjunto rectangular de esquinas de la ciudad por la que los robots pueden circular. Estas pueden ser de tres tipos:

- Área compartida (areaC): es el tipo de área por defecto y corresponde a una región de la ciudad de libre acceso, es decir, cualquier robot puede circular por ella.
- Área privada (areaP): una región de este tipo sólo permite que haya un robot en ella. El intento de un robot de ingresar en un área privada de otro, genera un error en tiempo de ejecución. Notar que las áreas privadas permiten un mecanismo de

exclusión mutua implícita entre robots.

- Área parcialmente compartida (areaPC): este tipo de regiones permiten el acceso de uno o varios robots, con la restricción de que deben haber sido autorizados previamente. Notar que las áreas parcialmente compartidas permiten un mecanismo de exclusión mutua selectiva entre robots.

Cada declaración de área comienza con un nombre, seguido de dos puntos y la palabra clave areaC, areaP o areaPC (para indicar su tipo) más cuatro parámetros. Éstos representan las coordenadas inferior izquierda y superior derecha que ocupará el área dentro de la ciudad. Cada tipo de área tiene asociado un color, tal como muestra en la Figura 3.

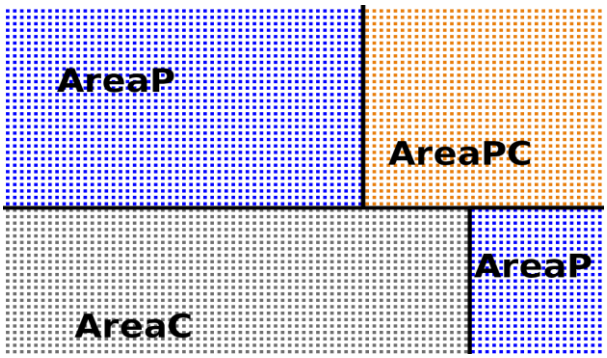


Fig. 3. Esquema de la ciudad con las definiciones de las distintas áreas.

El alcance y la visibilidad de las áreas corresponden a todo el programa, y deben ser asignadas a los robots sobre los que se quiera permitir su acceso antes de comenzar la ejecución de los mismos.

3.2 Declaración de robots

La declaración de robots comienza con la palabra clave **robots** y termina donde se encuentra la palabra clave **comenzar**.

Los robots tienen una estructura casi idéntica a la del programa principal o subprocesos, incluyendo encabezado, declaraciones y cuerpo.

El encabezado comienza con la palabra clave **robot** seguido de un nombre. Las declaraciones de procesos y variables locales siguen las mismas reglas que las declaraciones del programa principal, con la salvedad que no pueden declararse nuevas áreas o robots.

De esta manera, la creación de robots siempre es explicita. Una característica a destacar es que aun cuando se quiera utilizar el entorno para simular un ambiente como el del Visual Da Vinci, será necesario crear el único robot desde el programa principal.

El cuerpo de un robot es una secuencia de sentencias, delimitada por las palabras clave **comenzar** y **fin.** Estas sentencias se corresponden con las de Visual Da Vinci, y sirven para definir el comportamiento del robot en la ciudad.

La posibilidad de definir tipos de robots permite reutilizar el código para diferentes

robots que llevan a cabo el mismo comportamiento, teniendo en cuenta que se debe tener cuidado principalmente con el uso de ubicaciones absolutas, dado que los robots puede que compartan o no un área de la ciudad.

3.3 Cuerpo del programa principal

Desde el programa principal es necesario asignar los robots a la áreas que pertenecen y luego "arrancar" cada uno ellos mediante la directiva *Iniciar*, que requiere del nombre del robot y su ubicación inicial.
Esta sentencia requiere de comprobaciones en tiempos de compilación para que dos o más robots no intenten ocupar la misma esquina, teniendo en cuenta también a qué tipo de área pertenece la esquina involucrada. Para dar un ejemplo, un robot no puede ocupar una esquina que pertenece a un área exclusiva de otro robot.
A estas sentencias se agrega un nuevo subconjunto, denominado sentencias de concurrencia.

3.4 Manejo de colisiones

Conceptualmente los "recursos compartidos" en este modelo de entorno se pueden reducir al acceso a una esquina, donde puede haber objetos.
Evitar las "colisiones" en las esquinas es el problema básico de sincronización en el entorno CMRE.
Para esto el lenguaje cuenta con directivas que permiten bloquear y liberar el recurso:

- *bloquearEsquina* (BE): indica que el robot pide exclusión para la ocupación de una esquina (lo que a su vez le permite recoger o depositar objetos).
- *liberarEsquina* (LE): indica que el robot deja el recurso libre (la esquina ocupada).

3.5 Comunicación / Sincronización

Tal como se mencionó anteriormente, hay múltiples robots que trabajan en la ciudad. En muchos casos, deberán colaborar en la resolución de algún problema.
Esto requiere comunicación y sincronización. Se adopta un mecanismo explícito de pasaje de mensajes asincrónicos con dos directivas:

- *enviarMensaje* (EM): permite que un robot envíe un mensaje a otro (identificados por su nombre). Al enviar el mensaje, según el modelo asincrónico, el robot continúa con la siguiente instrucción secuencialmente sin esperar la recepción.
- *recibirMensaje* (RM): indica que un robot se quedará esperando hasta sincronizar con el envío de mensaje de otro. En la recepción se indica el nombre del robot del cual se espera el mensaje.

Para finalizar se eligieron dos problemas ampliamente utilizados en la enseñanza de los conceptos de programación concurrente y paralela.
En la Figura 4 se muestra el código correspondiente a un problema "master/worker" que utiliza Pasaje de Mensajes. Para esto se declaran 4 aéreas privadas junto a 4 robots (1,2,3 workers ,4 master) donde cada uno interactúa en su área privada juntado todas las flores que existen en ella, y al finalizar los robots 1, 2 y 3 envían sus resultados al 4 para que este los totalice e informe dicho total.

```
programa ejemplo1

procesos
  proceso avenida(ES f:numero)
  comenzar
    repetir 49
      {recorre la avenida y junta las flores
        acumulando en el parámetro f}
  fin

areas
  {definición de áreas}
  area1: areaP(1, 1, 50, 50)
  area2: areaP(1, 51, 50, 100)
  area3: areaP(51, 1, 100, 50)
  area4: areaP(51, 51, 100, 100)

robots
  {comportamiento de c/tipo de robot}
  robot worker
  variables
    f:numero
  comenzar
    f:=0
    repetir 49
      avenida(f)
      Pos(PosAv+1,PosCa-49)
    avenida(f)
    enviarMensaje(f, robot4)
  fin
```

```
robot master
  variables
    f:numero
    total:numero
  comenzar
    f:=0
    repetir 49
      avenida(f)
      Pos(PosAv+1,PosCa-49)
    avenida(f)
    total:=f
    recibirMensaje(f, robot1)
    total:=total+f
    recibirMensaje(f, robot2)
    total:=total+f
    recibirMensaje(f, robot3)
    total:=total+f
    informar(total)
  fin

variables
  {creación variables robots}
  robot1: worker
  robot2: worker
  robot3: worker
  robot4: master

comenzar
  {Asignación de áreas a cada robot}
  AsignarArea(robot1, area1)
  AsignarArea(robot2, area2)
  AsignarArea(robot3, area3)
  AsignarArea(robot4, area4)
  iniciar(robot1, 1, 1)
  iniciar(robot2, 1, 51)
  iniciar(robot1, 51, 1)
  iniciar(robot2, 51, 51)
fin
```

Fig. 4. Ejemplo de programa con Pasaje de Mensajes.

En la Figura 5 se muestra el código correspondiente a un problema que utiliza Memoria Compartida. Para esto se declara que toda la ciudad es compartida por 2 robots (1 y 2) donde deben coordinarse para trasladar de a una las flores de la esquina (1,1) hasta que la misma queda vacía. Esta coordinación debe darse para garantizar que los robots no estén en la misma esquina simultáneamente, y por consiguiente no tomen la misma flor. Cada vez que un proceso toma una flor la traslada a la esquina siguiente (diferente para cada robot).

```
programa ejemplo2
procesos
  proceso girar(E cant:numero)
  comenzar
    repetir cant
      derecha
  fin
  proceso depositarUnaFlor
  comenzar
    mover
    liberarEsquina(1,1)
    depositarFlor
  fin

areas
  {definición de áreas}
  area1: areaC(100, 100, 100, 100)

robots
  {comportamiento de c/tipo de robot}
  robot tipo1
  variables
    seguir:boolean
  comenzar
   seguir:=V
    girar(2)
    bloquearEsquina(1,1)
    mover
    si ~(HayFlorEnLaEsquina)
       seguir:=F
    mientras(seguir)
      tomarFlor
      girar(2)
      depositarUnaFlor
      girar(2)
      bloquearEsquina(1,1)
      mover
      si ~(HayFlorEnLaEsquina)
         seguir:=F
    girar(2)
    mover
    liberarEsquina(1,1)
  fin
```

```
  robot tipo2
  variables
    seguir:boolean
  comenzar
    seguir:=V
    girar(3)
    bloquearEsquina(1,1)
    mover
    si ~(HayFlorEnLaEsquina)
      seguir:=F
    mientras(seguir)
      tomarFlor
      girar(2)
      depositarUnaFlor
      girar(2)
      bloquearEsquina(1,1)
      mover
      si ~(HayFlorEnLaEsquina)
        seguir:=F
    girar(2)
    mover
    liberarEsquina(1,1)
  fin

variables
  {creación variables robots}
  robot1: tipo1
  robot2: tipo2

comenzar
  {Asignación de áreas a cada robot}
  AsignarArea(robot1, area1)
  AsignarArea(robot2, area1)
  iniciar(robot1, 1, 2)
  iniciar(robot2, 2, 1)
fin
```

Fig. 5. Ejemplo de programa en Memoria Compartida.

3.6 Estado del desarrollo actual

El entorno CMRE está totalmente desarrollado en Java y se está utilizando experimentalmente en la UNLP. Los robots físicos están en proceso de compra, aunque no introducen (en el estado actual del desarrollo) una complejidad adicional.

Las características elegidas de los robots físicos se relacionan con nuevas posibilidades que se abren para el entorno, tal como se indica en las líneas de trabajo futuras.

4 Conclusiones y Líneas de Trabajo Futuro

Se ha presentado un entorno para la enseñanza temprana de los conceptos de concurrencia y paralelismo, asociados con el empleo de robots virtuales y físicos en un entorno de programación interactivo y flexible.
Actualmente se está estudiando la generalización del empleo del CMRE en aplicaciones en las cuales los robots adquieren información en tiempo real y los algoritmos definidos en el entorno toman decisiones dinámicamente. Esto es de particular importancia para asignaturas relacionadas con Sistemas de Tiempo Real e incluso con Sistemas Inteligentes.

5 Bibliografía

[ACM04] ACM/IEEE-CS Joint Task Force on Computing Curricula. "Computer Engineering 2004: Curriculum Guidelines for Undergraduate Degree Programs in Computer Engineering". Report in the Computing Curricula Series. 2004.

[ACM08] ACM/IEEE-CS Joint Interim Review Task Force. "Computer Science Curriculum 2008: An Interim Revision of CS 2001". Report from the Interim Review Task Force. 2008.

[ACM13] ACM/IEEE-CS Joint Task Force on Computing Curricula. "Computer Science Curricula 2013". Report from the Task Force. 2013.

[ACM68] ACM Curriculum Committee on Computer Science. "Curriculum „68: Recommendations for the undergraduate program in computer science". Communications of the ACM, 11(3):151-197. 1968.

[ACM78] ACM Curriculum Committee on Computer Science. "Curriculum „78: Recommendations for the undergraduate program in computer science". Communications of the ACM, 22(3):147-166. 1979.

[ACM99] ACM Two-Year College Education Committee. "Guidelines for associate-degree and certificate programs to support computing in a networked environment". New York: The Association for Computing Machinery. 1999.

[AMD09] AMD. "Evolución de la tecnología de múltiple núcleo". http://multicore.amd.com/es-ES/AMD-Multi-Core/resources/Technology-Evolution. 2009.

[AND00] Andrews G. "Foundations of Multithreaded, Parallel, and Distributed Programming". Addison Wrsley, 2000.

[CAR03] Carr S., Mayo J., Shene C. "Threadmentor: a pedagogical tool for multithreaded programming". ACM Journal of Educational Resources, 3:1–30, 2003.

[DAS89] Dasgupta S. "Computer Architecture. A Moder Synthesis. Volume 2: Advanced Topics". Jhon Wilet & Sons. 1989.

[DEG12a] De Giusti A. E., Frati F. E., Leibovich F., Sánchez M., De Giusti L. C., Madoz M. C. "Concurrencia y Paralelismo en CS1: la utilización de un Lenguaje Visual orientado". Proceeding del VII Congreso de Tecnología en Educación y Educación en Tecnología. 2012

[DEG12b] De Giusti L. C., Frati F. E., Leibovich F., Sánchez M., Madoz M. C. "LMRE: Un entorno multiprocesador para la enseñanza de conceptos de concurrencia en un curso CS1".

Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología. Págs. 7 - 15. 2012.
[DEG12c] De Giusti A. E., Frati F. E., Sánchez M., De Giusti L. C. “LIDI Multi Robot Environment: Support software for concurrency learning in CS1”. Proceeding of IEEE International Conference on Collaboration Technologies and Systems. Pág. 294-298. 2012.
[DEG13] De Giusti A. E. “El cambio tecnológico como motor de la Investigación en Informática”. Conferencia inaugural del Workshop de Investigadores en Ciencia de la Computación (WICC2013). 2013.
[DIJ65] Dijkstra E. W. “Solution of a problem in concurrent programming control”. Communications of the ACM, 8(9):569, 1965.
[DIJ78] Dijkstra E. W. “Finding the Correctness Proof of a Concurrent Program”. In Program Construction, International Summer Schoo, Friedrich L. Bauer and Manfred Broy (Eds.). Springer-Verlag, 24-34, 1978.
[GEP06] Gepner P., Kowalik M.F. “Multi-Core Processors: New Way to Achieve High System Performance”. In: Proceeding of International Symposium on Parallel Computing in Electrical Engineering 2006 (PAR ELEC 2006). Pags. 9-13. 2006.
[GOL72] Goldstine H. H. “The Computer”. Princeton University Press, 1972.
[GPG] GPGPU. “General-Purpose Computation on Graphics Processing Units”. http://gpgpu.org.
[HAN77] Hansen P. B. “The Architecture of Concurrent Processes”. Prentice Hall, 1977.
[HOA78] Hoare C. “Communicating Sequential Processes”. Communications of the ACM, 21(8): 666-677, 1978.
[HOA85] Hoare C. “Communicating Sequential Processes”. Prentice Hall, 1985.
[HOO13] Hoonlor A., Szymanski B. K., Zaki M. J., Thompson J. “An Evolution of Computer Science Research”. Communications of the ACM. 2013.
[HWA84] Hwang K., Briggs F. A. “Computer Architecture and Parallel Processing”. McGraw Hill, 1984.
[HWA93] Hwang K. “Advanced Computer Architecture: Parallelism, Scalability, Programmability”. McGraw Hill, 1993.
[LEGO] "Lego Education". http://www.legoeducation.us/eng/characteristics/ProductLine~LEGO%20MINDSTORMS%20Education%20EV3.
[LEGOb]Lego. "LEGO Mindstorms EV3 Announced". http://brickextra.com/2013/01/10/lego-mindstorms-ev3-announced/
[MCC08] McCool M. “Scalable Programming Models for Massively Parallel Multicores”. Proceedings of the IEEE, 96(5): 816–831, 2008.
[OLS83] Olsen E. W., Whitehill S. B. “Ada for Programmers”. Prentice Hall, 1983.

Una propuesta para la incorporación de Cloud Computing en la currícula de Grado

Nelson Rodríguez[1], María Murazzo[2], Daniela Villafañe[3], Adriana Valenzuela[4], Adriana Martin[5], Susana Chavez[6]
Departamento e Instituto de Informática, Universidad Nacional de San Juan, Complejo Universitario Islas Malvinas, Rivadavia, San Juan, Argentina
nelson@iinfo.unsj.edu.ar[1], maritemurazzo@gmail.com[2], villafane.unsj@hotmail.com[3], franciscaadriana.valenzuela@gmail.com[4], arianamartinsj@gmail.com[5],schavez@iinfo.unsj.edu.ar[6]

Abstract. Cloud Computing es un modelo de provisión de recursos que está transformando los modos tradicionales de cómo las empresas utilizan y adquieren los recursos de Tecnología de la Información. La expansión de Cloud ha determinado la necesidad de formación de recursos humanos especializados. La mayoría de las universidades han resuelto este problema con especializaciones, maestrias o cursos ad-hoc. Pero ninguna ha revisado sus planes de estudio, para incluir adecuadamente los conocimientos básicos de Cloud en el grado si afectar considerablemente la currìcula. Este trabajo sugiere cuales son estos temas fundamentales asociados a Cloud Computing y propone la profundidad con que deben ser abordados para cualquiera de las carreras cuyo contenidos curriculares son aprobados por la resolución 786/2009.

Keywords: Cloud Computing. Curricula Informatics, Cloud Computing Curricula, Computer and Information Science Education, Curriculum

1 Introduction

Cloud Computing (CC) es un paradigma que está cambiando en gran parte la forma en que se hacen los negocios por Internet. Sin embargo existen distintas interpretaciones y enfoques de que es y no es Cloud Computing. Al existir tantas definiciones y algunas diferentes entre sí, utilizar el término puede resultar engañoso. Algunos usuarios creen que con solo utilizar algún servicio como e.mail en Internet es suficiente para decir que están en Cloud, otros especialistas son más puristas y consideran que si no están soportando los servicios en una verdadera plataforma Cloud son servicios provistos a través de Internet, pero no Cloud Computing. Con el objetivo de consensuar el concepto en la industria, la revista Cloud Computing Journal reunión a 20 expertos [1] y publicó las distintas definiciones las cuales presentaban coincidencias y diferencias.

El nuevo modelo de negocio y prestación de servicios actual, es el Cloud Computing. Los recursos que provee Cloud responden a las necesidades de las empresas u organizaciones que quieran dar uso a las mismas.

Los servicios que ofrece Cloud se pueden agrupar en las categorías. Así, Cloud Computing permite "alquilar" infraestructura hardware en la red (IaaS, infraestructura como servicio), utilizar plataformas colaborativas y herramientas de desarrollo disponibles en Cloud (PaaS, plataforma como servicio) o directamente consumir aplicaciones de software ofrecidas por el proveedor de servicios o pertenecientes a la propia empresa que permitirán mejorar su organización interna y ofrecer servicios online avanzados a sus clientes (SaaS, Software como servicio).

Desde el punto de vista académico se han realizado varias iniciativas, ya sea por medio de cursos específicos como: el curso CS309A, Cloud Computing de la universidad de Stanford[2]; o CS5412: Cloud Computing (Spring 2012) Universidad de Cornell[3], entre otros. Algunas empresas además ofrecen certificaciones y otros varios cursos sobre tecnologías específicas para Cloud [4]. Además se pueden encontrar especializaciones y maestrías, las cuales por supuesto incluyen 2 o más cursos como la maestría en Cloud Computing de la Universidad de Newcastle [5].

También se debe destacar las primeras jornadas de Cloud Computing realizadas en Argentina, llevadas a cabo en la Facultad de Informática de la UNLP [6] y también el curso de Postgrado dictado en la misma facultad por Dr. Xoan Pardo [7].

En todos los casos los esfuerzos son válidos, pero no resuelve el problema de que los contenidos básicos sean trasmitidos en el grado.

Teniendo en cuenta que la informática y en particular Cloud Computing sufre una constante evolución, es necesario revisar los planes de estudio periódicamente para incorporar estos nuevos contenidos y plantear los objetivos actitudinales.

Cuando se incorporaron alumnos a los proyectos de investigación, se comprobó que muchos conceptos de Cloud Computing no eran conceptualizados por los alumnos debido a que no formaban parte de los contenidos de las materias obligatorias (en algunos casos se impartían a un nivel muy introductorio en materias optativas), y por ende no formaban parte de los conocimientos que tiene un egresado.

A partir de ahí surgió el interés de discutir cuales eran los temas considerados más importantes, además ya se había realizado un trabajo anterior pero solo sobre los temas de Arquitectura Sistemas Operativos y Redes que fue publicado en TE&ET [8].

Los planes de estudios vigentes en las carreras de nuestra universidad se basan en lo propuesto por la red UNCI. Por ello, el punto de referencia para analizar cuales contenidos la industria propone como necesarios deberían ser aquellos no presentes en la resolución 786/2009 [9].

Dicha resolución define los contenidos curriculares básicos, carga horaria mínima, criterios de intensidad de la formación práctica y estándares de acreditación referidos a la carreras de Licenciatura en Ciencias de la Computación, Licenciatura en Sistemas/Sistemas de Información/Análisis de Sistemas, Licenciatura en informática, Ingeniería en Computación e Ingeniería en Sistemas de Información/Informática,

Para llegar a una propuesta válida, se debería analizar a todas las entidades o personas interesadas en producir modificaciones a los planes de estudio.

La industria necesita especialistas en determinadas áreas, los organismos como la IEEE y ACM, Red UNCI proponen currículas, las universidades proponen cursos y

los egresados a través de su experiencia profesional aportan lo suyo. Es evidente que un análisis pormenorizado de lo que han documentado todos los interesados.

Por lo tanto el punto de partida es la resolución 786, se tuvieron en cuenta los temarios de cursos de la industria y las universidades.

También se ha tenido en cuenta las sugerencias que están expresadas en los borradores de la futura propuesta CS 2013 (que modifica las áreas de conocimiento de las carreras) [10].

2 Cloud Computing

Los avances en IT han permitido que supercomputadoras y clusters soporten millones de operaciones concurrentes. El paralelismo está soportado por la comunicación y coordinación, estas dos actividades han sido transformadas dramáticamente. Además las comunicaciones de alta velocidad quasi-ubicuas no solo posibilitan que los data centers sean reubicados sino también que las computaciones sean movidas a facilidades centralizadas que ejecutan economía de escala y permite que enormes cantidades de datos sean agrupados y organizados para soportar tareas de decisión de usuarios por todo el mundo. Los gobiernos, laboratorios de investigación y empresas necesitan simular fenómenos complejos, similarmente Google, Facebook, Microsoft y otras CSP (Proveedoras de Servicio de Cloud) necesitan procesar grandes cantidades de datos operados en masivos datacenters “cloud”.

El paralelismo masivo, la comunicación ultraràpida y la centralización masiva serán fundamentales para la toma de decisión humana. Las computaciones que serán usadas para predecir el tiempo, indexar la Web, recomendar películas, restoranes y hoteles, sugerir conexiones sociales y más, son distribuidas sobre cientos de procesadores y dependen de colecciones de datos, a veces de millones de fuentes repartidas por todo el mundo [11].

La importancia de Cloud Computing puede verse reflejada en varias estadísticas publicadas sobre inversión y sus perspectivas. Por ejemplo en un estudio realizado por Gartner, predicen el tamaño del mercado de Cloud Computing podría alcanzar los 150 mil millones de dólares en 2013.

Por otro lado Mimecast realizó un estudio estadístico y encontró que 7 de cada 10 empresas que utilizan servicios Cloud moverá nuevas aplicaciones al mismo. Sólo es el 70%, porque varias respondieron que no quieren “poner todos los huevos en una sola canasta". Esto significa que algunas empresas todavía se muestran escépticas acerca de mudarse completamente a Cloud.

Gartner también predijo que el 60% de las cargas de trabajo de servidor se virtualizarán en 2014, debido a la cantidad de beneficios que obtiene a cambio, como es reducir la compra de hardware, la huella de carbono y los costos de energía. Esta es una gran manera de ahorrar dinero en el largo plazo [12].

El National Institute of Standards and Technology ha presentado una de las definiciones de Cloud más clara y comprensible. La define como un modelo que habilita acceso a red ubicuo, conveniente, bajo demanda para compartir un conjunto de recursos configurable, que pueden ser rápidamente provistos y liberados con mínimo esfuerzo o interacción del proveedor de servicios. Distingue las

características de Cloud, el modelo de entrega y los métodos de desarrollo. Resalta así, los cinco (5) aspectos claves de cloud computing: auto servicio bajo demanda, acceso a la red ubicua, un conjunto de recursos independiente de la ubicación, rápida elasticidad y servicio a la medida [13].

Cloud Computing no es un desarrollo revolucionario reciente, sino es el resultado de la evolución de varias tecnologías. Conceptos precursores son: utility computing, computación bajo demanda, computación elástica o grid computing [14].

Se puede pensar a Cloud Computing como un modelo de aprovisionamiento de recursos IT que potencia la prestación de servicios IT y servicios de negocio, facilitando la operativa del usuario final y del prestador del servicio. La característica básica de este modelo es que los recursos y servicios informáticos, tales como infraestructura, plataforma y aplicaciones, son ofrecidos y consumidos como servicios a través de la Internet sin que los usuarios tengan que tener ningún conocimiento de lo que sucede detrás.

Cloud Computing es un esquema que a veces se expresa como XaaS o EaaS, para significar Everything as a Service. Usualmente se divide a Cloud Computing en las siguientes capas: Software como Servicio (SaaS), Plataforma como Servicio (PaaS) e Infraestructura como Servicio (IaaS).

Investigaciones recientes de IDC muestra los ingresos públicos en todo el mundo de TI, donde los servicios de Cloud superaron los $ 16 mil millones en 2009 y se prevé llegar a 55,5 mil millones dólares en 2014, lo que representa una tasa compuesta de crecimiento anual del 27,4%. Esta rápida tasa de crecimiento es más de cinco veces el crecimiento previsto para los productos tradicionales de TI (5%).

Frank Gens, Senior VP y Analista jefe en IDC dice: “Un reciente estudio entre Ejecutivos de IT, CIOs y los colegas en las líneas de negocio muestra que la Cloud Computing está ‘cruzando el abismo’ y entrando en un período de amplía adopción. Más aún, la crisis económica amplificará la adopción de Cloud. Este modelo ofrece una manera más barata para que el negocio use y adquiera tecnología. Esta ventaja es verdaderamente importante para los pequeños y medianos negocios, un sector que será clave en cualquier plan de recuperación [15].

Esta fuerte presencia de Cloud Computing en el mercado está cambiando el perfil del profesional de IT. Al respecto la Debra Littlejohn Shinder, comenta que aspectos que eran complementarios ahora son centrales, a tal fin describe las 10 áreas claves para especialistas de IT en los próximos años, donde figura Cloud Computing en primer lugar [16].

Cloud Computing generará una década de investigación en virtualización, computación distribuida, utility computing, redes, servicios de software y servicios Web. Implica una arquitectura orientada a servicio, de gran flexibilidad con reducido costo de propiedad, con servicios bajo demanda y muchas otras cosas [17].

Además se cita en varias predicciones que hacen algunas consultoras como Garnet, que en octubre de 2012 expuso que entre las 10 principales tecnologías consideradas clave por la consultora están: Personal Cloud, The Internet of Things, Hybrid IT and Cloud Computing [18].

3 Qué contenidos involucra Cloud Computing

Se puede considerar a Cloud Computing como “la multidisciplinariedad dentro de la disciplina”, porque si se tiene en cuenta la informática, Cloud Computing involucra conceptos de Sistemas Distribuidos, conectividad y Sistemas Operativos (ARSO), bases de datos NSQL, metodologías de desarrollo específicas para Cloud y lenguajes de programación también específicos, inclusive en la actualidad se están utilizando servicios de Cloud para aplicaciones para HPC.

Según Charles Border "Cloud Computing es una nueva palabra de moda para un grupo de viejas tecnologías que han sido integradas para crear un sistema que es más que la suma de todas las partes” [19].

Por lo tanto, surge como un desafío determinar cuáles son estos contenidos mínimos y cómo poder incluirlos en la curricula con el menor impacto posible.

Vale la pena resaltar, que el objetivo de este trabajo es el expuesto en el párrafo anterior, y no en sugerir una carrera de especialista en Cloud, que es ese caso debería incluir varios contenidos adicionales, o sea, cuales son los contenidos para un Licenciado en Sistemas o en Ciencias de la Computación o en Informática o en alguna de las otras terminales que deben impartirse y que establezcan las bases para el conocimiento de Cloud Computing a nivel de grado.

Teniendo en cuenta cursos, capacitaciones de empresas y publicaciones, se pueden determinar que Cloud involucra grandes temas, los cuales son: virtualización, arquitectura cloud (IaaS, PaaS, SaaS y sus variantes), Data Centers, Big Data, Seguridad.

Cada una de las áreas que abarca o se relaciona con CC presenta muchos contenidos para desarrollar por ejemplo XaaS, o sea cualquier cosa como servicio y cuando se definió los temas a incluir se tuvo en cuenta también las tecnologías que hacen posible CC. Estas tecnologías son necesarias que sean estudiadas antes de tratar CC porque en ellas se fundamenta el paradigma, y si no se han tratado con suficiente profundidad es posible que el crédito horario de la presente propuesta deba ser ampliado.

En la Resol.786 figuran los siguientes tópicos como base para CC: algoritmos concurrentes, distribuidos y paralelos, algoritmos y lenguajes, concurrencia y paralelismo, Concepto de arquitectura basadas en servicios, Seguridad en Sistemas Distribuidos, Arquitecturas de Multicomputadoras y Computación orientada a Redes.

Pero además se deben tener en cuenta las tecnologías habilitantes son: tecnologías de multicomputadoras y multithreading, Computación HPC, Redes, Datacenter, Virtualizaciòn, SOA, Modelos de Programación Distribuida y paralela.

La importancia de todas las tecnologías de base para CC se puede justificar en el hecho de que las CS2013 agregar como nueva área a la Computación Paralela y Distribuida (al ser un área incluye más de una materia) [10], y también muchas empresas además de las tradicionales en CC han comenzado a ofrecer servicios de Cloud como las Telco, que han evolucionado desde servicios de conectividad, luego ISP y ahora se han transformado en CSP.

4 Antecedentes

Aunque mucho se ha escrito acerca de integrar nuevas tecnologías en la curricula, muy poco ha sido escrito acerca de la integración de Cloud Computing en los planes de estudio.

Un año antes del inició un proyecto de investigación sobre CC (en 2019) se comenzaron a realizar investigaciones y publicaciones sobre este paradigma, se hicieron presentaciones en la WICC [21,22,23,24,25], CACIC [26], COMTEL (Perù) [27], RUEDA [28], SABTIC[29,30] y JCC [31,32].

En una encuesta realizada a egresados de carreras de informática, se determinó que entre los temas sugeridos como importantes aparecen virtualización y Cloud Computing, temas que están incluidos en todos los cursos de CC [8].

En dicha encuesta, se valoró de la siguiente forma: 5 – Muy necesario 4- Necesario 3-Necesite a veces conocerlo 2- Pocas veces necesite conocerlo 1-No me hizo falta nunca. Como resultado arrojó que CC llegó a una ponderación de 2,75, programación de alta performance 2,80, programación paralela 2,45, Cluster 2,5 y virtualización 3,4 sobre 5 puntos.

Por otro lado se toma en cuenta la resolución 786 que es la que está vigente para las curriculas de carreras ligadas a la informática. La propuesta de este trabajo no es la de generar una materia nueva o un grupo de materias sobre Cloud Computing, sino de como agregar los mínimos contenidos sin modificar considerablemente la currícula. Si bien miembros de la red UNCI están trabajando en modificaciones a los planes de estudio y muchos de los contenidos de CC se han incorporado a estas modificaciones a partir de lo que se conoce como la iniciativa informático 2020, estos cambios pueden verse reflejados en el mejor de los casos en un par de años y solo para los alumnos ingresantes.

En la materia Sistemas Distribuidos, correspondiente al 5to año de la Licenciatura en Ciencias de la Computación el la UNSJ. Se realizaron durante 2012 y 2013 experiencias educativas al incorporar una introducción a CC en 2012 y se profundizaron algunos temas y se incorporó una práctica sencilla sobre un PaaS en 2013. Esto permitió junto con otras experiencias como Conferencias y Tutoriales expuestos en distintos ámbitos, calcular el tiempo necesario para incorporar estos contenidos a la currícula [33,34,35].

Para la parte práctica se decidió trabajar con Google App Engine, porque se presentaban algunas dificultades administrativas para el uso de Amazon Web Service y Azure en el Cloud. Por otro lado es conveniente realizar las mismas en el ambiente que demande menor tiempo en conocerse (ya sea por el ambiente o el lenguaje), teniendo en cuenta el plan de estudios o conocimientos previos de los estudiantes, por ejemplo puede ser que se conozca mejor Python, Java o C#, y por lo tanto se elegirá el ambiente más adecuado.

Aunque la experiencia resultó exitosa, teniendo en cuenta que los alumnos comprendieron la temática, quedan temas para profundizar que deberían impartirse en otras materias a lo largo del plan de estudios, como por ejemplo virtualización en sistemas operativos y Big Data en base de datos o materias relacionadas.

Para elaborar la propuesta se tuvo en cuenta demás dos trabajos realizados que tratan la problemática de incluir CC en la currìcula, como son el presentado por Charles Border en el SIGSE'13 [19] y el de James Lawler en el EDSIG 2011 [36]. El

primer trabajo trata sobre los Fundamentos y tecnologías que hacen posible CC, mientras que el segundo desarrolla una propuesta similar a la presente pero tomando en cuenta curriculas IS 2009 publicada por la IEEE y ACM, pero generada a partir de BMP (Business Process Management), o sea con un enfoque a Sistemas e IS, y no general para las diferentes carreras.

CC va a tener un impacto similar o inclusive superior al de seguridad como fue expuesto en la propuesta de curricula 2013 [10], debido a que el impacto de CC es sobre todo el plan de estudios y por lo tanto no es conveniente agregar una materia sobre Cloud sino distribuir los contenidos en las distintas materias relacionadas.

5 La Propuesta

El objetivo de la propuesta es introducir las bases del paradigma emergente de Cloud Computing para que sean dictados en las materias de grado sin modificar sustancialmente los créditos horarios de las asignaturas.

El alumno debería conocer:
Los fundamentos de Cloud Computing
Las tecnologías de apoyo
Entender las limitaciones, ventajas y desventajas,
Técnicas de creación y uso del Cloud
Los fundamentos de MapReduce como modelo de programación

Es conveniente que la mayoría de los conceptos se impartan en los años superiores (4to y 5to año), debido a que el tiempo necesario para las distintas actividades (teorías y prácticas) se puedan realizar en el menor tiempo posible y no se incremente el crédito horario de las materias, aunque la incorporación de contenido nuevo siempre obliga a realizar determinados ajustes.

Los contenidos propuestos son:
Conceptos introductorios (1 hora)
Clasificación de servicios: SaaS, PaaS, IaaS (1 hora)
Modelos de despliegue: público, privado e híbrido (1 hora)
Virtualización y Data Centers (2 horas)
Clusters y Arquitecturas de HPC (2 o 3 horas)
Base de Datos NoSQL y Big Data (1, 2 o 3 horas, no se contempla práctica)
MapReduce (1 hora)
Programación del Cloud y Ambientes de Software (1 hora, si se pone énfasis en un solo ambiente, sino puede ser hasta 3 horas)
Ambientes de Software Emergentes: Open Source, Eucalyptus y Nimbus (1 hora)
Ciclo de vida y Metodología para Cloud Computing (2 horas)
Prácticas que pueden ser sobre PaaS o SaaS (4 horas o más)
Aspectos Legales de Cloud Computing, fundamentalmente SLA para Cloud (Service Level Agreement) (1 hora, que debe ser impartido en el espacio de materias del área Aspectos Profesionales y Sociales)

En el mejor de los casos son solo 18 horas que pueden ser reubicadas en distintas materias, pero en el peor de los casos serán 23 horas.

Cabe hacer notar que virtualización es un tema que se encuentra dentro de la resolución 786, por lo que dicho tema se está impartiendo en las carreras que abarca dicha resolución, pero la hora que se agrega es para enseñar los contenidos de virtualización de middleware e hipervisores.

Quedan además varios temas por profundizar o conceptualizar, que serían objeto de un curso específico (o más de un curso) y no como parte de la currícula para un alumno de grado como son: Eficiencia en energía para centro de datos, métricas de performance y escalabilidad, métricas de tolerancia a falla y disponibilidad, Hadoop a un nivel avanzado, HPC sobre Cloud, Seguridad específica en Cloud, otros casos de servicios Cloud como Desktop como servicio, o Database como servicio, Monitoreo como Servicio,

Por otro lado CC puede ser considerado como un modelo de entrega de SaaS, PaaS e IaaS, pero además Database como servicio, Información como servicio, Integración como servicio, Administración como servicio, Plataforma como servicio, Proceso como servicio, Seguridad como servicio, Almacenamiento como servicio y Testing como Servicio.

6 Conclusiones

Cloud Computing no solo es una buzzword, es el modelo de cómo se construirán gran parte de las aplicaciones del futuro. La inversiones que están realizando las empresas en Cloud son millonarias y los planes de estudio deben reflejar lo que está pasando en la industria, porque nuestro alumnos se insertarán en ese mercado laboral, por supuesto sin descuidar los fundamentos y la formación que debe tener un profesional universitario.

Para cada una de las carreras de las distintas terminales: Licenciatura en Ciencias de la Computación, Licenciatura en Sistemas de Información, Licenciatura en informática, Ingeniería en Computación e Ingeniería en Sistemas de Información, se deberán plantear los contenidos específicos de CC, por lo tanto queda a partir de este trabajo mucho para discutir.

La propuesta no incluye un crédito horario razonable para los conocimientos básicos de CC y no demanda grandes adecuaciones para llevarlo a cabo.

Desde el punto de vista educativo CC favorece la integración de diversos contenidos como paradigmas de computación (Web Services, data Centers, Utility Computing, Grid Computing, P2P e Internet de las Cosas), Modelos de Programación Distribuida y Paralela, y atributos y capacidades deseadas (Ubicuidad, Confiabilidad, Escalabilidad, QoS, SLA y Aspectos legales y consideraciones Sociales), lo que impactará fuertemente en la formación del alumno.

Cloud Computing irá cambiando conforme aparezcan nuevas investigaciones y desarrollos, y por supuesto estos cambios también impactarán en la curricula.

References

1. Cloud Computing Journal: Twenty-One Experts Define Cloud Computing. http://cloudcomputing.sys-con.com (2008)
2. Chou T.: CS309A Cloud Computing. Universidad de Stanford http://scpd.stanford.edu/search/publicCourseSearchDetails.do?method=load&courseId=11815
3. Birman K. CS5412 Cloud Computing. Universidad de Cornell. http://www.cs.cornell.edu/courses/cs5412/2012sp/
4. Private Cloud certification Solutions Expert, Microsoft, http://www.microsoft.com/learning/en-us/private-cloud-certification.aspx
5. MSc Cloud Computing, Universidad de Newcastle, http://www.ncl.ac.uk/computing/study/postgrad/taught/5056/
6. I Jornadas de Cloud Computing, III-LIDI, Facultad de Informática, UNLP, http://jcc2013.info.unlp.edu.ar/
7. Pardo X. (UDC): Cloud Computing, Curso de Postgrado Doctorado en Ciencias Informáticas, http://postgrado.info.unlp.edu.ar/Cursos/Cursos_2013/06-2013_Cloud_Computing.pdf
8. Rodríguez N., Murazzo M., Villafañe D.: Cuáles son los conocimientos de ARSO (Arquitectura, Redes y Sistemas Operativos) que la industria considera importantes, VII Congreso TEyET 2012.
9. Ministerio de Educación: Resolución 786/2009, 26/5/2009, disponible en: http://redunci.info.unlp.edu.ar/docs/BoletinOficial_Resolucion_786-2009.pdf
10. The Joint Task Force on Computing Curricula Association for Computing Machinery
IEEE-Computer Society: Computer Science Curricula 2013,Ironman Draft (Version 1.0) February 2013, http://redunci.info.unlp.edu.ar/docs/cs2013-ironman-v1.0.pdf.
11. 13. Hwang K., Fox G., Dongarra J.: Distributed and Cloud Computing from Parallel Processing to the Internet of Things (eds.) Morgan Kauffmann (2012)
12. Williams N., Marketing Coordinator, WebSan Solutions Inc., a Canadian Certified Microsoft Partner: 3 Interesting Cloud Computing Statistics, Junio 2013, http://www.erpsoftwareblog.com/2013/06/3-interesting-cloud-computing-statistics/
13. Mell P., Grance T.: NIST: Definition of Cloud Computing, Special Publication 800-145, (2011)
14. Zhu J., Fang X., Guo Z., Hua Niu M., Cao F., Yue S., Liu Q.: IBM Cloud Computing Powering a Smarter Planet, Libro Cloud Computing, Volumen 5931/2009, Páginas 621-625.(2009)
15. IDC :IDC Finds Cloud Computing. Entering Period of Accelerating Adoption and Poised to Capture IT Spending Growth Over the Next Five Years, http://www.idc.com/getdoc.jsp?containerId=prUS21480708
16. Littlejohn Shinder D. MVP,: 10 hot areas of expertise for IT specialists, TechRepublic, Feb (2011).
17. Youk M..:Cloud Computing – Issues,Research and Implementations, Journal of Computing and Information Technology, CIT 16, 2008, 4, 235–246. doi:10.2498/cit.1001391.
18. Gartner Identifies the Top 10 Strategic Technology Trends for 2013. Analysts Examine Top Industry Trends at Gartner Symposium/ITxpo, October 21-25 in Orlando, Press Release, http://www.gartner.com/newsroom/id/2209615 (2012).
19. Border C.: Cloud Computing in the Curriculum: Fundamental and Enabling Technologies. SIGCSE'13 ACM (2013).
20. Murazzo M., Rodríguez N.: Mobile Cloud Computing, XII WICC 2010, Calafate, (2010).
21. Rodríguez N., Chávez S., Martín A., Murazzo M., Valenzuela A.: Interoperabilidad en Cloud Computing. XIII WICC, Rosario (2011).

22. Chávez S., Martin A., Rodríguez N., Murazzo M., Valenzuela A.: Metodología AGIL para el desarrollo SaaS, XIV WICC, Mayo 2012. Posadas, Misiones (2012).
23. Rodríguez N., Valenzuela A., Chavez S., Martin A., Murazzo M., Villafañe D.: Ambiente de desarrollo para lengua de señas basado en cloud, XIV WICC, Mayo 2012. Posadas, Misiones (2012).
24. Martín A., Chávez S., Rodríguez N., Valenzuela A., Murazzo M.: Bases de Datos NoSql en Cloud Computing, XV WICC Abril 2013.Paranà Entre Ríos (2013).
25. Rodríguez N., Murazzo M., Villafañe D., Alves M., Medel D.: Integración de Computación Heterogénea con Hadoop para Cloud Computing, XV WICC Abril 2013.Paranà Entre Rìos (2013).
26. Murazzo Maria, Millán Flavia, Rodríguez Nelson, Segura Daniela, Villafañe Daniela (Oct. 2010).”Desarrollo de aplicaciones para cloud computing”. CACIC 2010. Morón. (2010).
27.Murazzo María, Rodríguez Nelson: Una propuesta para el desarrollo de aplicaciones para mobile cloud computing. Congreso Internacional de Computación y Telecomunicaciones – COMTEL 2010, Lima, Perú. Oct. (2010).
28.Millán F., Murazzo M., Rodríguez N. (2010): Plataformas educativas implementadas con mobile cloud computing, V Seminario Internacional “De legados y Horizontes para el Siglo XXI”, organizado por RUEDA, Tandil. Sep. (2010).
29. Rodrìguez N., Murazzo M., di Sciacio C. : Integración de Computación móvil con Cloud Computing, 1º Seminario Argentina Brasil de Tecnologías de la Información y la Computación, Rosario noviembre (2011).
30. Rodrìguez N., Villafañe D., Murazzo M., Gallardo D., Tarrachano G.: Integración de las capas SaaS / Paas del Cloud en la tecnología Google, 2º Seminario Argentina Brasil de Tecnologías de la Información y la Computación – Tres de Maio Brasil (2012).
31. Rodríguez N., Murazzo M., Chávez S., Valenzuela A., Martín A., Villafañe D.: Aspectos claves para el desarrollo de aplicaciones para Mobile Cloud Computing, JCC 2013. La Plata (2013)
32. Murazzo M., Rodríguez N., Villafañe D., González F.: Perspectivas en el análisis de grandes volúmenes de datos en el Cloud. JCC 2013. La Plata (2013)
33. Rodríguez, Murazzo, Ene: Cloud Computing, Workshop de investigadores en
Ciencias de la Computación y Sistemas de Información. San Juan. Nov. (2009).
34. Murazzo M. Segura D., Villafañe D.: Cloud Computing con Windows Azure, 2° Jornadas de Actualización Informatica. San Juan junio, (2010).
35. Rodríguez N., Villafañe D.: Cloud Computing (conferencia invitada). 2da Jornadas organizadas por CASETIC (Cámara de Empresas de Software). San Juan. Oct. (2010).
36. Lawler J. :Cloud Computing in the Curricula of Schools of Computer Science and Information Systems, Information Systems Education Journal (ISEDJ) (2011).

Propuesta de una metodología para una rápida enseñanza de circuitos lógicos y de su integración en una UCP en carreras de Informática

Mario Carlos Ginzburg
Profesor Titular de "Sistemas de Computación I y II"
en la Universidad Abierta Interamericana
mario.ginzburg@uai.edu.ar

Resumen. Los programas de Arquitectura de Computadoras y similares de carreras de Informática son extensos, y en general presentan una sola unidad dedicada a circuitos lógicos, debiéndose formar en limitadas clases a los alumnos en los conocimientos básicos. Esta problemática trae aparejado un debate acerca de los contenidos mínimos a enseñar acorde con la formación buscada, que en esta propuesta se plantean y desarrollan. Opcionalmente, conforme al perfil y nivel de cada carrera, los circuitos de la UCP tratados aisladamente pueden integrarse en un modelo de UCP didáctico, también desarrollable en pocas clases. Esta plataforma no requiere conocimientos de electrónica, y también puede emplearse en estudios terciarios de Informática y en carreras de Electrónica como introducción a los circuitos digitales. La presente metodología didáctica se ha concretado con excelentes resultados en ingenierías informáticas de la FIBA (3er. año) y Facultad de Tecnología de la UAI (1er. año) desde el 2002 al presente.

Palabras claves: circuitos lógicos, enseñanza didáctica rápida

1. Introducción

Cuando alumnos de carreras de Informática de distintas partes del país solicitan equivalencias de estudios, se observa al leer los programas de Arquitectura de Computadoras o asignatura similares, que en general se dispone de pocas clases para desarrollar desde “cero” los contenidos dedicados a los circuitos básicos que conforman la UCP (compuertas, decodificadores, UAL, multiplexores, flip flops, registros y buses). Estos circuitos en general se enseñan desconectados entre sí, sin conformar una UCP. En las carreras de Electrónica se tiene una asignatura como Técnicas Digitales dedicada por completo a los circuitos combinacionales y secuenciales.
Por las limitaciones de tiempo señaladas se requiere que la enseñanza de esos circuitos sea sintética y conceptual, a la par que didáctica. Para lograr esto es necesario discutir primero qué contenidos son propios de Informática, como se plantea en el ítem 2 de este trabajo
Si además, por el perfil y exigencias del nivel de estudios de una carrera de Informática, dichos circuitos digitales deben integrarse en un modelo didáctico simple de UCP, su tiempo de enseñanza se vuelve doblemente crítico si se quiere desarrollar normalmente los restantes contenidos de Arquitectura de Computadoras.
Tal modelo de UCP construido con compuertas y flip flops que forman parte de sus bloques funcionales, debe permitir, sin usar electrónica, visualizar y comprender en profundidad en cada ciclo del pedido y ejecución de instrucciones, entre otras cosas, el papel temporal y circuital de los MHz y de las líneas de control sobre: los registros, la UAL, la memoria y los caminos de datos (“data paths”). Asimismo debe poder ilustrar acerca de cómo interactúan hardware y software, y cómo es que en la decodificación el código de operación de una instrucción determina el valor que tendrán las líneas de control para que el hardware la ejecute y pida la siguiente. Además debe ser factible en dicho modelo ejecutar instrucciones de salto en las que intervienen los flags, para comprender cómo la máquina decide entre dos alternativas, y poner en claro que la UCP no tiene inteligencia propia para concretar en cada ciclo sus acciones. Igualmente debe ser útil para establecer diferencias entre CISC y RISC.
A fin de poder desarrollar con rapidez este modelo, es indispensable que previamente los alumnos hayan sistematizado en un modelo sin circuitos lógicos, sólo con “cajas negras” de la UC, la UAL y los registros interconectadas por buses, los movimientos y cambios que ocurren en un computador cuando se piden y ejecutan las instrucciones. De esta forma se sigue el orden de lo general a lo particular, para luego volver a lo general.

2. Problemática abordada

Conforme al objetivo buscado de que en las carreras de Informática la enseñanza de los circuitos lógicos sea sintética, conceptual, didáctica, y sean mínimos sus tiempos de enseñanza, primero es indispensable definir qué aspectos de la enseñanza de circuitos son necesarios y suficientes en estas carreras, para luego desarrollar un proceso de enseñanza-aprendizaje apropiado y didáctico que cualquier docente puede aplicar.

La presente metodología es fruto de sucesivos perfeccionamientos didácticos practicados durante más de diez años de enseñanza de los contenidos aquí planteados. En tal sentido, el autor como director de cátedras, llevó a cabo observaciones empíricas siguiendo la evolución en el aprendizaje a través de cuestionarios, exámenes parciales y finales de centenares de alumnos de las asignaturas "Estructura del Computador ", de 3er. año de Ingeniería Informática de la FIBA (2000-2006), y Sistemas de Computación II de 1er. año la Facultad de Tecnología Informática de la UAI (del 2004 al presente). Debe consignarse que la mayoría de los alumnos provienen de establecimientos secundarios no técnicos, sin conocimientos de electrónica o electricidad.
A fin de no perder tiempo en dibujos por parte del docente o de los alumnos, esta metodología requiere proyectar imágenes, que pueden formar parte de un texto, o un apunte con un atlas de figuras, al cual los alumnos pueden remitirse para sistematizar y repasar todos los temas tratados en clase.

2.1 Enseñanza tradicional con aspectos relacionados con carreras de electrónica

Habitualmente, en parte por que los primeros profesores de materias del tipo Arquitectura de Computadoras en facultades de Informática fueron ingenieros electrónicos, el aprendizaje de los circuitos lógicos en general ha sido muy semejante al que tiene lugar en los primeros tramos de la asignatura Técnicas Digitales de las carreras de Ingeniería Electrónica. Un objetivo propio en estas carreras es que el futuro ingeniero esté capacitado para proyectar subsistemas electrónicos digitales.
Asimismo, aún hoy, con la influencia de Técnicas Digitales, además de métodos algebraicos de minimización y transformaciones circuitales, se enseñan detalladamente un conjunto de flip flops asincrónicos y sincrónicos, en sus distintos funcionamientos (R-S, J-K, T, etc.), siendo que, como se desarrolla en el ítem 3, es factible a partir de un multiplexor de 2 entradas construir de manera simple un flip flop "D" sincrónico, con el cual se pueden implementar los registros y secuenciadores que necesita un modelo simple de UCP.
Esta concepción en la enseñanza de los temas básicos también se apoya en la bibliografía clásica para Arquitectura de Computadoras, con textos que contemplan tanto el funcionamiento como el diseño de computadores, objetivo que en principio no está presente en el perfil del ingeniero en Informática o Sistemas.
En Stallings [1] se enseñan los temas antes mencionados, y además se agrega simplificación por el método de Quine-Mc Cluskey y contadores sincrónicos. En Tanenbaum [2] si bien aparece la enseñanza tradicional de este tema, no se dan métodos de simplificación. En Murdocca - Heuring [3], en Hamacher - Vranesic - Zaky [4], y en Alcalde - Portillo - García Merayo [5], por citar algunos, contienen bien detalladas metodologías apropiadas para carreras de ingeniería electrónica. Inclusive en algunos textos aparecen transistores y hojas con datos eléctricos de compuertas.
Una limitación que presenta en general esta enseñanza, es que si bien los alumnos tienen un primer conocimiento del funcionamiento de circuitos lógicos tratados aisladamente, no ven su funcionalidad formando parte de una UCP.

2.2 Objetivos propios, planteo didáctico y herramientas para la enseñanza de circuitos lógicos en carreras de Informática

En la formación del ingeniero en Informática es secundario el proyecto de circuitos lógicos, ya sea con mínima cantidad de compuertas, o con compuertas de un solo tipo, o para integrarlos en un chip de un computador. Esto es propio de un ingeniero electrónico, que necesita una enseñanza intensa del álgebra de Boole y de la síntesis con chips de muy alta integración circuital, amén de conocimientos profundos sobre física electrónica, electricidad y electrónica.
En las carreras de informática interesa en primer lugar *el funcionamiento* de los circuitos citados en el item 1 que componen una UCP, y si se pretende un nivel de enseñanza superior, dichos circuitos pueden interconectarse a fin de conformar una UCP, sin apelar a la electrónica.
El Álgebra de Boole debe ser una herramienta formativa en Informática, en el sentido de que los alumnos en primer lugar puedan entrever isomorfismos entre las funciones Or, And y negación y estructuras lógicas del lenguaje Assembler y de lenguajes de alto nivel. Así, existen a nivel de máquina instrucciones en bajo nivel que ordenan operaciones And, Or, negación, X-Or usadas en programas; e inversamente, mediante una estructura IF puede describirse la tabla de funcionamiento de una compuerta.
Didácticamente la enseñanza se simplifica, es más comprensible y se reducen tiempos, si los circuitos se construyen sólo con compuertas And, Or e inversores, como ocurre en la metodología planteada.
Además es necesario conocer los símbolos de la lógica clásica para comprender formalmente las operaciones lógicas que realiza la UAL, o escribir en forma compacta mediante sumas de productos un comportamiento circuital And-Or, con el que puede expresarse verbalmente cualquier tabla de funcionamiento mediante las conectivas "o" e "y". Resulta así un correlato conceptual y didáctico entre el álgebra de Boole y el álgebra de proposiciones.
Esto es la base para que los alumnos de Informática puedan construir un circuito And-Or suma de minitérminos a partir de una tabla de funcionamiento que debe cumplir, desarrollando de esta forma en ellos aptitudes generales para

proyectar. Así, a partir de la tabla de un sumador completo, se sintetiza un circuito que cumpla con ella, para poder construir luego con 4 u 8 de éstos el sumador/restador de una UAL y los flags que ella genera.
Si bien una expresión booleana tiene la relevancia cognitiva y formativa de expresar formalmente con pocos símbolos una tabla o un comportamiento circuital, didácticamente es discutible si al inicio de la enseñanza de los circuitos lógicos se deben desarrollar forzosamente métodos algebraicos para simplificar usando Karnaugh, o para pasar de un circuito a otro equivalente mediante De Morgan. Estos desarrollos producen discontinuidades didácticas en el proceso de enseñanza-aprendizaje del funcionamiento de los circuitos lógicos. Esto se manifiesta en que al abordar el tema de la UAL, o siguientes, muchas veces los alumnos, por el tiempo transcurrido, no tienen presentes temas básicos vistos al comienzo, por lo que hay que volver a repasarlos.
Por otra parte, una enseñanza analítica de circuitos basada exclusivamente en expresiones booleanas puede tornarse demasiado abstracta o árida para muchos alumnos.
En este sentido, si se quiere verificar algebraicamente que una expresión booleana se corresponde con una tabla, el alumno debe reemplazar en forma abstracta valores binarios en esa expresión. Este procedimiento puede volverse algo mecánico y atemporal, alejando al alumno de una primera apreciación empírica y directa del funcionamiento de un circuito And-Or corriente.
En cambio, como se ejemplificará, una simple inspección visual inmediata de cada And con inversores que conforma dicho circuito And-Or, determinará directamente qué combinación binaria hace valer uno la salida de la And, y en consecuencia también la salida de la Or. De este modo se va construyendo conceptualmente la tabla buscada.
Si en lugar de una expresión booleana, se parte de una tabla de funcionamiento, siempre se tiene una visión totalizante del comportamiento que debe proveer cualquier implementación circuital que se deduzca de ella, y de hecho proporciona más información que un plano con compuertas que simboliza un circuito digital que cumple con esa tabla, o que la expresión algebraica que sintetiza su respuesta.
Consecuentemente, a los fines didácticos, en la metodología presentada se dará más preeminencia a las tablas que a sus expresiones algebraicas.
Por otra parte en la asignatura Matemática Discreta de la carrera de Informática es corriente enseñar el álgebra de Boole, incluidas las formas normales y la minimización, ocurriendo a veces superposiciones en la enseñanza de temas.

En relación con la enseñanza de flip flops, no es imprescindible partir de los asincrónicos, y es suficiente desarrollar el "D" sincrónico, con sus dos estados fácilmente recordables: uno cuando la salida copia la entrada (con Ck=1) y el otro cuando retiene (con Ck=0). Con este flip flop se construirán los registros de la UCP y las celdas de memoria.
De esta forma, si con cuidado se prescinde de contenidos propios de Ingeniería Electrónica, es factible reducir las horas de clases que ello implica, y se pueden desarrollar en un par de clases los circuitos de la UCP y memoria.
En el corto tiempo que insume este aprendizaje tipo "mecano", se van construyendo uno tras otro, circuitos combinacionales y secuenciales de complejidad progresiva, usando sólo compuertas And, Or e inversores. Ello redunda en que sin saltos temáticos los alumnos puedan integrar y sistematizar más fácilmente los aspectos comunes y diferentes del funcionamiento de los mismos.
A posteriori es factible integrar estos circuitos para conformar una UCP básica completa, que también puede desarrollarse en un reducido número de clases.

3. Esquema con aspectos básicos de la plataforma propuesta

En el inicio del proceso de enseñanza-aprendizaje presentado se procura que el alumno pueda establecer correspondencias y equivalencias conceptuales y visuales de significados simbólicos entre una tabla, un circuito And-Or que la verifica y una suma de productos minitérminos que lo expresa.
Al respecto, a partir de las tablas de las compuertas And y Or se verifica y pone de relieve que su denominación está relacionada con el hecho de que la relación entre entradas y salida se puede enunciar con proposiciones lógicas con las conectivas "o" e "y", a la par que se definen la simbología circuital de cada compuerta y su expresión algebraica, poniendo de relieve que son tres formas distintas de simbolizar lo mismo.
A continuación en cada una de las entradas de una And se conecta o no un inversor (conexión que abreviaremos Inv-And), se determina la tabla de este conjunto y se verifica visualmente en el conexionado que la salida de la And vale 1 para una sola combinación presente en las entradas del conjunto Inv-And. Didácticamente conviene poner de relieve por simple examen visual, que para un conjunto Inv-And existe una única combinación que puede hacer que todas las entradas de la And valgan 1 para que su salida valga 1. Como ilustra el Inv-And más alto de la figura 1 su salida vale 1 sólo si en sus entradas se tiene Ai = 0 "y" Bi =1 "y" Ci-1 =1 que producirán 1 "y" 1 "y" 1 en las entradas de la And.
Los alumnos así internalizan la correspondencia: combinación binaria <==> Inv-And que la detecta unívocamente.
En lo que sigue, didácticamente usaremos en forma modular y visual conjuntos Inv-And para detectar combinaciones con dos usos complementarios: a) dado un Inv-And construido, determinar visualmente qué combinación da valor 1 a su

salida; y b) dada una combinación a detectar, construir por simple inspección visual un Inv-And que la identifique mediante el valor 1 de su salida.

Tabla 1

A_i	B_i	C_{i-1}	C_i	S_i
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

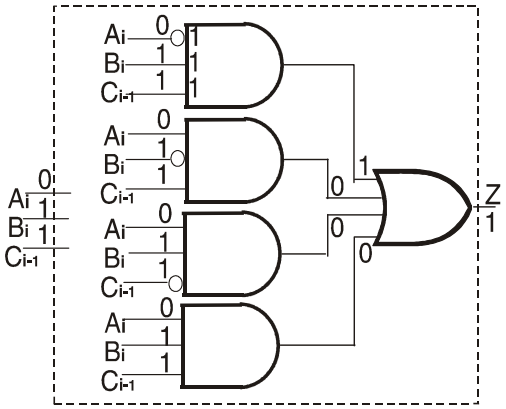


figura 1

Luego de las verificaciones con Inv-And, en las que fundamentalmente interviene la vista, ya pueden desarrollarse circuitos decodificadores de 2 y 3 entradas, basados en estos módulos Inv-And detectores de combinaciones, lo cual también sirve para que los alumnos afiancen conocimientos.
Para el decodificador de 3 entradas los alumnos deberán construir 8 módulos Inv-And para identificar del 000 al 111, y hacer el cableado necesario para que todos reciban juntos la combinación presente en las entradas del decodificador. Además verificarán que sólo tendrá valor 1 la salida del decodificador correspondiente al Inv-And construido para detectar la combinación presente en las entradas del decodificador.
En Informática es importante vincular el funcionamiento de un decodificador con el carácter random de cualquier memoria electrónica, planteando que en éstas el número binario de n bits que recibe el decodificador es la dirección de la celda a acceder, y que cada una de sus 2^n salidas termina en una celda, para permitir el acceso a ella cuando es direccionada.

Seguidamente puede analizarse un circuito Inv-And-Or de 3 entradas, en el que varios Inv-And de igual número de entradas concurren a una Or como en la figura 1 con el objeto de hallar por inspección visual la tabla que el circuito verifica, supuesta desconocida.
A tal fin primero los alumnos deben reconocer que algunos de los Inv-And son los mismos que se han tratado en el decodificador. O sea que los 4 Inv-And del circuito dado conforman un decodificador que no detecta 8 sino sólo 4 combinaciones distintas (011, 101, 110, 111), determinables visualmente por la ubicación de los inversores en cada Inv-And.
La salida de la Or toma valor 1 si cualquier salida de un Inv-And vale 1. Ello ocurrirá si la salida del Inv-And superior detecta, como muestra la figura 1 (por la ubicación de su inversor) que en las entradas del circuito se ha recibido 011 (Ai=0 "y" Bi=1 "y" Ci-1=1); "o" si el Inv-And siguiente detecta que se ha recibido 101 (Ai=1 "y" Bi=0 "y" Ci-1=1); "o" si el anteúltimo Inv-And detecta que se ha recibido 110; "o" si el último Inv-And detecta que se ha recibido 111.
Así se obtienen las 4 combinaciones 011, 101, 110 y 111 a las que les corresponden los 4 "unos" de la tabla 1. Como cada una de las restantes combinaciones no puede generar salida 1 en ninguno de los 4 Inv-And, ellas producirán 4 ceros en las entradas de la Or, o sea salida cero en ella.
Con esta técnica, partiendo de un circuito Inv-And-Or, con igual número de entradas en cada And, con sólo examinarlo visualmente, se puede determinar la tabla que cumple.
A cualquier circuito Inv-And-Or como el analizado se le puede hacer corresponder un enunciado del álgebra proposicional como el anterior usando las conectivas "o" e "y", lo cual permite al alumno tener una comprensión más clara de su funcionamiento lógico.
Este tipo de enunciados se pueden formalizar en el álgebra de Boole mediante sumas de productos minitérminos, que se corresponden circuitalmente a estructura Inv-And-Or con compuertas And de igual número de entradas.
Si se quiere también puede hallarse la suma de productos minitérminos propia del circuito con las correspondencias simbólicas "+" y "." para las conectivas "o" e "y" respectivamente, y con las convenciones simbólicas que se aplican para escribir minitérminos. Así el funcionamiento anterior puede expresarse:

$$Ci = \overline{Ai}.Bi.Ci\text{-}1 + Ai.\overline{Bi}.Ci\text{-}1 + Ai.Bi.\overline{Ci\text{-}1} + Ai.Bi.Ci\text{-}1 .$$

No es obligatorio tratar este tema al comienzo de la enseñanza. Tampoco se plantea hallar la tabla buscada reemplazando en esa suma de minitérminos cada combinación, de 000 a 111 en este caso, siendo que ella se ha obtenido concretamente observando directa y simplemente las entradas de los Inv-And, para así determinar las combinaciones (tantas como Inv-And distintos existan) para las cuales la columna de salida de la tabla vale 1.

Habiendo los alumnos asimilado la metodología visual de análisis del comportamiento de un circuito, puede plantearse el procedimiento inverso: a partir de una tabla como la tabla 1 (que en este ejemplo debe cumplir la salida Ci de un semisumador), construir paso a paso el circuito Inv-And-Or (suma de minitérminos) que la verifica, el cual aparece terminado en la figura 1.
El tipo de circuito buscado ya fue analizado por los alumnos, y al igual que cualquiera a sintetizar, será del tipo Inv-And-Or, de estructura y funcionamiento lógico que se corresponde con un enunciado que emplea las conectivas "o" e "y".
Procediendo de forma inversa al análisis, a partir de la tabla se determina que en este caso son 4 la cantidad de combinaciones a detectar que deben generar el valor 1 en la salida de la Or final. Este número también establece que serán necesarios 4 módulos detectores Inv-And, cuyas salidas irán a una Or que así tendrá 4 entradas (figura 1). La ubicación de los inversores en cada uno de los 4 Inv-And sigue la misma técnica usada en la construcción del decodificador, o sea depende de la posición de los ceros de cada combinación a detectar que debe hacer valer 1 a la salida.
Así se llega al circuito de la figura 1, y luego con 4 semisumadores se puede construir un sumador/restador de 4 bits, que forma parte de una UAL.
Empleando esta metodología de síntesis, a partir de la tabla de funcionamiento de que se trate, los alumnos están capacitados para construir el correspondiente circuito combinacional Inv-And-Or.
Con la presente propuesta, desde el inicio de la enseñanza de circuitos lógicos es factible ahorrar muchas horas de clase sin desviar la atención del alumno en temas como métodos algebraicos de simplificación de funciones, transformación de expresiones booleanas en Nand-Nand, u otros.

La enseñanza de flip flops parte de un circuito selector (multiplexor) de 2 entradas (figura 2) cuya entrada de selección S oficiará de reloj (Ck) en el flip flop sincrónico "D" que resulta simplemente de realimentar la salida Q en la entrada inferior del selector (figura 3).
Los alumnos comprenden en primer lugar que de este modo el cable de realimentación obliga a que el valor que tiene la salida Q del selector sea siempre el mismo que presenta la entrada inferior, o sea que el propio circuito da valor a una de sus entradas, por lo que así ésta no recibe valor 0/1 del exterior.

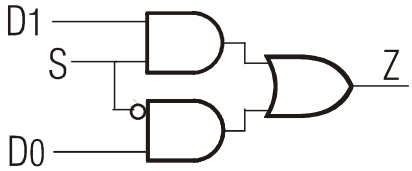


Figura 2

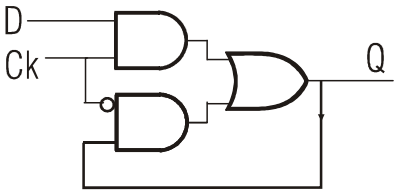


figura 3

Los alumnos comprenden en primer lugar que de este modo el cable de realimentación obliga a que el valor que tiene la salida Q del selector sea siempre el mismo que presenta la entrada inferior, o sea que el propio circuito da valor a una de sus entradas, por lo que así ella no recibe valor 0/1 del exterior.
A continuación resulta claro, con los valores indicados (figura 4), que mientras sea Ck = 1 se selecciona la entrada D vinculada al exterior, por lo que la salida Q "copiará" el valor 0/1 que D recibe: si D mantiene su valor así lo hará Q, pero si D cambia también lo hará Q, tantas veces como lo haga D, sin posibilidad de mantener su valor.
Esta situación (como ilustra el dibujo superior de la figura 4) se simboliza con fines didácticos y mnemónicos con la caja del flip flop en la cual un cable virtual une la salida Q con la entrada D, siendo que en un cable conductor un extremo tiene el mismo valor (voltaje) que el otro.

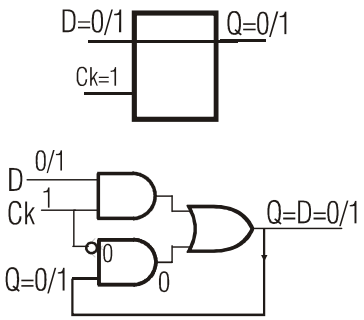


figura 4

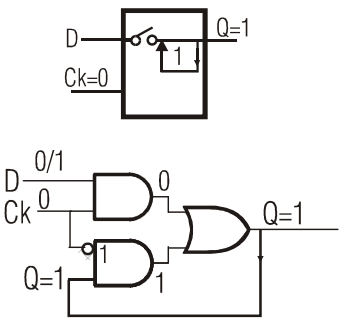


figura 5

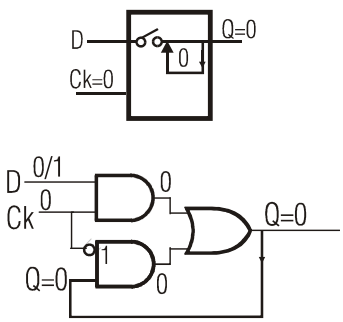


figura 6

Si luego se hace Ck = 0 (figuras 5 y 6) se selecciona la entrada inferior, que continuamente tiene el valor 0/1 de la salida, por estar conectadas ambas por el cable que las une. Entonces, el valor que tenga Q en el instante en que es Ck = 0 (que es

el mismo que tenía la entrada D en ese momento), será el de la entrada inferior. Conforme al funcionamiento del selector este valor debe aparecer en la salida Q, por lo que el valor que tenía Q en el instante en que Ck = 0 no podrá modificarse mientras sea Ck = 0, dado que en la entrada inferior permanentemente está ese mismo valor de Q (figuras 5 y 6).
Así el valor de Q se autorepite, se automantiene, quedando retenido, memorizado, aunque el valor de la entrada D cambie al valor contrario, dado que por no estar D seleccionada, su valor no puede cambiar la salida Q.
Como ilustran los dibujos superiores de las figuras 5 y 6 se simboliza mnemónicamente en cada caso esta situación de retención de un cero o uno mientras sea Ck = 0, con la salida Q autoinyectándose el valor que tiene, y la entrada D aislada de la salida, sin poder intervenir en el valor de Q.
Para cambiar el valor que mantiene la salida Q, se debe hacer Ck = 1, con lo cual se volverá a seleccionar la entrada D para que su valor pase a la salida, al mismo tiempo que se deja de seleccionar la otra entrada, para que la salida no se copie a sí misma. Si bien el valor realimentado de la salida Q está siempre presente en la entrada inferior debido al cable que las une, sólo se repite en Q cuando esta entrada está seleccionada por ser Ck = 0.
Con este tipo de flip flop sincrónico resulta más sencillo de entender a los alumnos cómo un circuito puede retener un bit. Asimismo es el único que necesitan conocer en la presente plataforma de enseñanza. Su funcionamiento se fija y sistematiza mediante los diagramas temporales correspondientes.
No es imprescindible desarrollar los flip flops asincrónicos como el R-S, ni partir de éste para enseñar flip flops.

Mediante flip flops "D" sincrónicos puede conformarse un registro, como se indica en la figura 7.
Conforme a la experiencia docente acumulada al respecto, puede estimarse que los temas hasta acá planteados pueden desarrollarse en 2 clases de 5 hs. cátedra.

4. Aplicación del método propuesto a la enseñanza de una porción de la UCP

El ejemplo siguiente apunta por un lado a poner de manifiesto las ventajas didácticas en la enseñanza que resultan de visualizar los flips flops (M-E) que conforman los registros mediante los mnemónicos de sus 2 estados, con Ck = 1 y Ck = 0 antes simbolizados. Por otro lado permite apreciar cómo los alumnos pueden comenzar a incursionar progresivamente en un modelo de UCP, para ver de qué forma concreta los MHz generados por un cristal actúan en el hardware configurando ciclos, y cómo las líneas de control de la UC determinan el ciclo en que un registro debe cambiar.
Si bien el ejemplo de la figura 7 ilustra una porción limitada de una UCP, con la presente metodología didáctica puede construirse una UCP básica completa, que contenga: la UC con líneas de control cuyos valores ella genera en cada ciclo, la UAL, los caminos de datos ("data paths"), los registros básicos, el decodificador y una matriz de conexionado. Un modelo pedagógico de esta UCP que data del año 2007 se desarrolla en un cuadernillo del autor [6] destinado a los alumnos, que debe actualizarse en función de nuevos planteos didácticos planteados en este trabajo.
El registro típico de la figura 7 está conformado por flip flops sincrónicos D "maestro-esclavo" cuyos instantes de cambio están sincronizados por los pulsos (designados Ck) recibidos por todos los flip flops simultáneamente, siendo que estos pulsos llegan sin invertir a las entradas Ck de todos los esclavos (E), e invertidos a las entradas Ck de los maestros (M).
En el conjunto dibujado el registro es representativo del Puntero de Instrucciones de la UCP, al cual en uno de los ciclos de los MHz al valor (0110) que contienen sus E se le quiere sumar 0010, para que en ese ciclo los E pasen a guardar la dirección (1000) de la próxima instrucción a localizar en memoria.
La estructura y comportamiento de este registro son iguales a los de cualquier otro registro de la UCP vinculado a los caminos de datos ("data paths"). En el inicio del ciclo que corresponda se debe poder aportar una copia del contenido del registro involucrado a donde ordene la instrucción en curso, y si es necesario, en ese mismo ciclo se debe poder reemplazar dicho contenido por otro nuevo, cuando el cristal genera el pulso con que termina ese ciclo. Esto sólo es factible si los flip flops "D" sincrónicos son "maestro-esclavo" (M-E), o sea compuestos por dos de estos flip flops simples.
Como aparece en la figura 7, en un ciclo determinado una compuerta And genera un pulso reloj que permitirá, en su flanco ascendente, que los M retengan el nuevo contenido (1000), que será copiado por los E, que lo guardarán en el otro flanco. Para ello una entrada de dicha And recibe continuamente los pulsos de los MHz que genera un cristal, y la otra está conectada a una línea de control (LC) de la UC, cuya circuitería hace que esta LC tenga valor 1 durante todo este ciclo en que supuestamente debe cambiar el presente registro. Por lo tanto durante ese ciclo la salida de la And generará una réplica de la forma de onda que produce el cristal de los MHz, la cual así llegará a los flip flops del registro como señal Ck. Durante los ciclos en que la LC tenga valor 0, la salida de la And valdrá 0, por lo que también será Ck = 0 en los E, que así estarán en estado de retención, permitiendo que el registro almacene.
Más en detalle, en el inicio del ciclo en que LC = 1 y hasta el flanco ascendente del pulso de ese ciclo, como el cristal de los MHz genera 0 volts, la salida de la And valdrá 0, lo cual determina que en los E sea Ck=0, y en los M sea /Ck=1. Entonces, conforme al funcionamiento de estos flip flops, en ese lapso los E seguirán reteniendo el contenido (0110) que tenían, como en sus cajas simboliza cada salida reinyectando su valor, y estos cables transmitirán 0110 al sumador, que le suma 0010 resultando 1000 en sus salidas. Éstas van a las entradas D de los maestros, que por recibir Ck=1 se comportan como cables que copian 1000 a sus salidas Q que llegan a las entradas D de los esclavos.

Esta situación persiste hasta que la salida de la And valga 1 por ser LC=1 y por generar valor 1 el cristal de los MHz. Entonces, en los M será /Ck=0, por lo que pasarán a retener 1000 que era el último valor que tenían sus salidas Q en ese instante, como se simboliza ahora con sus salidas reinyectando su valor. Dado que los E reciben Ck=1, se comportarán como cables cuyas salidas Q copian el valor 1000 que reciben de las salidas de los M que están reteniendo. Por lo tanto, las salidas de los E, que son las salidas del registro, enviarán al sumador el valor 1000 que si bien no está retenido en los E, permanece invariable mientras los E retengan. El sumador adicionará 0010 y generará 1010, pero este nuevo valor no pasará a las salidas del registro, pues los M están reteniendo 1000, por lo que no permiten que 1010 pase a los E.
Al inicio del siguiente ciclo y en los sucesivos se supone que LC=0, por lo que la salida de la And será 0 y nuevamente en los E será Ck=0 y /Ck=1 en los M, repitiéndose la situación esquematizada en la figura 7 superior. Por lo tanto en dicho inicio los E pasarán a retener 1000, que era el último valor que tenían sus salidas Q en ese instante del inicio del siguiente ciclo. A su vez en los E será /Ck=1, por lo que sus salidas copiarán nuevamente el nuevo valor 1010 de las salidas del sumador, que ahora no puede pasar a los E dado que están en estado de retención hasta que no sea otra vez LC=1.

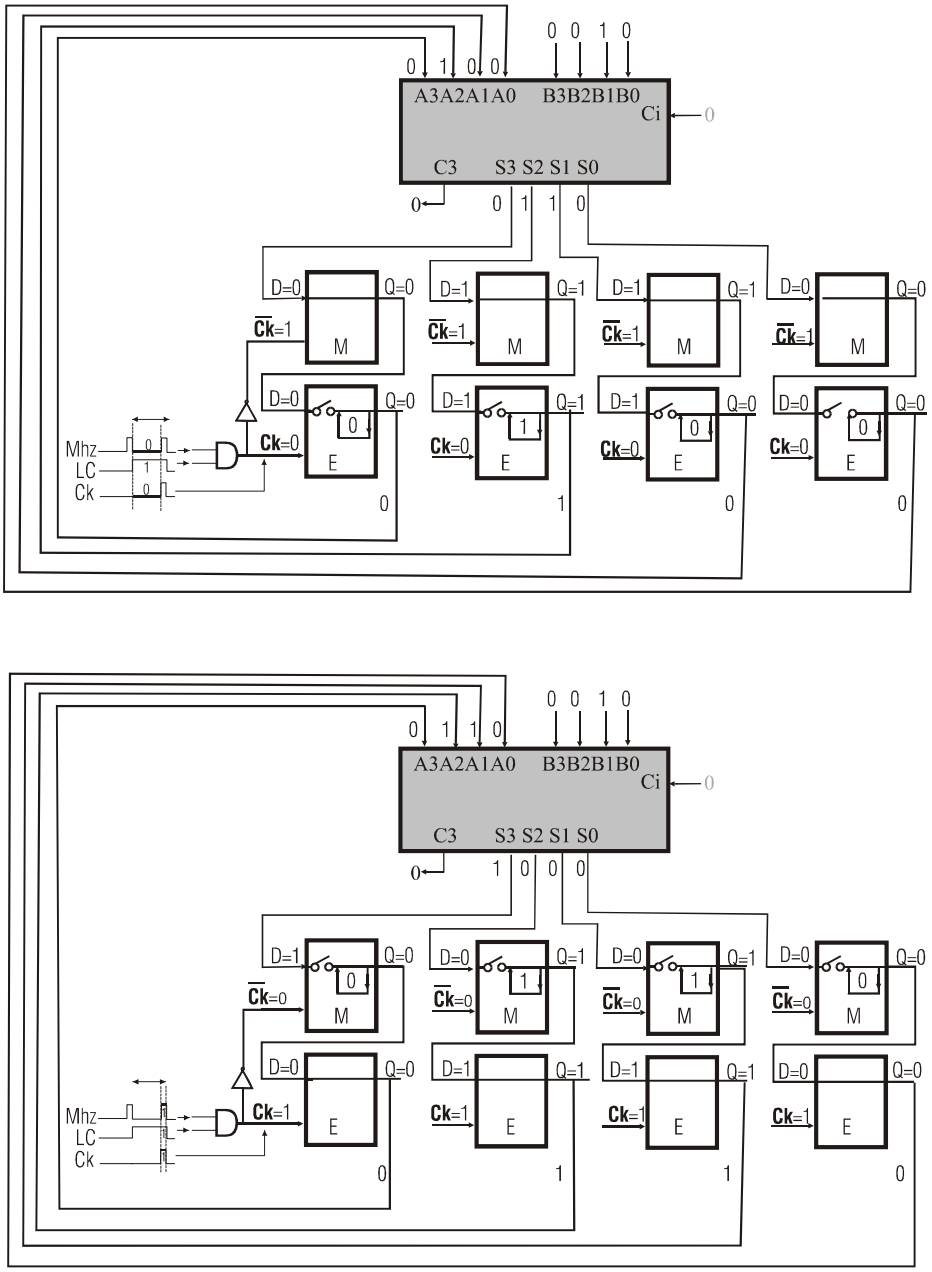


figura 7

De esta forma los alumnos, con la ayuda de los dos esquemas mnemónicos de los dos estados del flip flop, pueden visualizar más claramente este proceso que ocurre en un solo ciclo en el cual el registro primero aportó 0110 al sumador, y luego pasó a almacenar el nuevo valor 1000 que el sumador generó en un primer tramo del ciclo. Si bien en lo que resta del ciclo los E transmiten 1000 al sumador, el nuevo valor 1010 que pasa a generar el sumador no puede llegar a los E pues en ese lapso los M retienen 1000. Este ejemplo también sirve para que los alumnos entiendan la necesidad de que cada flip flop sea M-E. De no existir los maestros que al retener 1000 no "dejan pasar" el 1010, el sumador volvería a sumar 0010 n veces mientras sea Ck = 1.

5. Implementación de la metodología propuesta y del modelo

Como se indicó, el presente desarrollo didáctico ha sido puesto en práctica con excelentes resultados y perfeccionado continuamente en las asignaturas "Estructura del Computador", de 3er. año de Ingeniería Informática de la FIBA (2000-2006), y Sistemas de Computación II de 1er. año la Facultad de Tecnología Informática de la UAI (de 2004 al presente). También ha sido implementado sin problemas por los profesores adjuntos de Sistemas de Computación II, Ing. Enrique Douce e Ing. Ricardo Martín, del 2010 al presente.
Alumnos provenientes de escuelas comerciales manifestaron que nunca pensaron que podrían comprender tan en profundidad el funcionamiento de un procesador. También ha permitido avanzar más rápidamente en el dictado de otros temas como control microprogramado, RISC y CISC y el lenguaje Assembler. Asimismo los alumnos en los exámenes finales mostraron una comprensión más cabal del tema.

6. Conclusiones

Es factible en 2 clases de 5 hs. cátedra desarrollar con la metodología propuesta la enseñanza del funcionamiento de los circuitos lógicos básicos, y en 3 clases adicionales construir con ellos un modelo didáctico simple de UCP de características RISC.
Para los circuitos se parte de la capacidad de establecer correspondencias visuales sencillas e intuitivas entre combinaciones binarias y compuertas And con inversores en sus entradas, de modo que la salida de cada And valga 1 sólo para la combinación que se desea detectar con ella.
En forma didácticamente gradual, cada nuevo nivel alcanzado engloba los anteriores y permite pasar al siguiente, hasta llegar a los circuitos concretos básicos de una UCP.
El modelo simple de UCP permite que el alumno deje de tener una comprensión abstracta acerca de muchos detalles importantes que interesan, especialmente en relación con la temporización de las señales internas de la CPU durante la ejecución de las instrucciones, dado que el modelo le proporciona una forma concreta de visualizar en el espacio y tiempo el interior de una UCP. Como todos los modelos de este tipo ayuda entre otras cosas: a entender por qué una máquina no piensa; cómo una instrucción se ejecuta en pasos que ya están predeterminados por el hardware (construido por el hombre) en función del código de operación; a comprender el papel de los pulsos reloj y el de las señales de control en el manejo de los caminos de datos ("data path"); a visualizar los movimientos internos que tienen que ocurrir en dichos pasos; a constatar cómo los circuitos combinacionales transforman el código de operación en señales de control. De este modo, el alumno se sentirá capacitado para abordar, analizar y comparar modelos con pipe line y modelos de CPU reales.

7 Bibliografia

1. Stallings, W.: Computer Organization and Architecture. 9/E, Prentice Hall (2013).
2. Tanenbaum, A.: Structural Computer Organitation, 6/E, Prentice Hall (2012)
3. Murdocca M., Heuring V.; Principios de Arquitectura de Computadoras. Prentice Hall –Pearson Education (2002)
4. Alcalde E., Portillo J., Garcia Merayo F.: Arquitectura de Ordenadores, McGraw-Hill (1999)
5. Hamacher V, Vranesic Z, Zaky S. Computer Organization, McGraw-Hill (1996)
6. Ginzburg M.: De la Compuerta al Computador (2007), Biblioteca Técnica Superior (2007)

FUN: una herramienta didáctica para la derivación de programas funcionales

Araceli Acosta[1], Renato Cherini[1], Alejandro Gadea[1], Emmanuel Gunther[1], Leticia Losano[2], and Miguel Pagano[2]

[1] FaMAF - Univ. Nacional de Córdoba
{aacosta,cherini,gadea,gunther}@famaf.unc.edu.ar
[2] FaMAF y CONICET
{losano,pagano}@famaf.unc.edu.ar

1. Introducción

Existen diversas razones que nos llevan a reflexionar sobre las problemáticas del aprendizaje de la programación en contextos universitarios y de la formación universitaria de los profesionales de la programación. Por un lado, en el presente contexto donde el sector TIC tiene una incidencia cada vez mayor sobre el PBI nos debemos preguntar sobre las capacidades de la industria del software para brindar soluciones de calidad y proveer garantías de la corrección de los productos; en particular teniendo en cuenta la creciente importancia de métodos formales en la industria [6]. En este sentido es importante reflexionar sobre cómo la currícula de las carreras de computación ayuda a desarrollar las habilidades necesarias que permitan a sus estudiantes y egresados comprender las bases teóricas y utilizar las herramientas prácticas que permiten producir software confiable.

Por otro lado, la obligatoriedad de la educación secundaria da lugar a una población de ingresantes a las carreras de computación más heterogénea que en otras épocas, con distintos recorridos escolares, en particular en lo que respecta a las habilidades lógico-matemáticas. Esto nos presenta el desafío de lograr una mayor permanencia de los estudiantes en las carreras manteniendo el nivel de excelencia y calidad. La incorporación de tecnologías de la información a la enseñanza de los conceptos básicos de lógica y programación puede ser de gran utilidad para este propósito.

En este trabajo se describe la utilización de una herramienta desarrollada para la enseñanza de lógica y programación. En la sección 2 se presentan el contexto de utilización de la herramienta y se aborda la perspectiva didáctica; en la sección 3 se describen los conceptos básicos a enseñar con esta herramienta. En la sección 4 se describe la utilización de la herramienta con ejemplos. En la sección 5 cerramos con las conclusiones de nuestro trabajo.

2. Una perspectiva para enseñar programación

El primer curso de programación en la carrera de la Lic. en Cs. de la Computación de Fa.M.A.F es “Introducción a los algoritmos” y se dicta en el primer

semestre de primer año. El objetivo de este curso es introducir la programación en pequeña escala como la transformación de especificaciones formales en programas ejecutables; las habilidades que se espera los estudiantes adquieran son la capacidad de modelar problemas formalmente, el uso de la lógica como herramienta para razonar sobre y probar la corrección de programas. El dictado de este curso tiene una interesante experiencia acumulada durante diez años y fruto de ello es el libro [2], utilizado como material principal durante el semestre. Esta experiencia también se traduce en ciertas miradas reflexivas sobre el curso.

Desde 2007 algunos investigadores han desarrollado investigaciones focalizadas en este curso [5, 3] que permitieron comprender mejor los procesos de aprendizaje involucrados en el mismo y delinear algunas de las dificultades experimentadas por los estudiantes. A partir de entrevistas a estudiantes, en [3] se rescatan algunas percepciones de los estudiantes: (I) una importante discontinuidad entre los contenidos y las formas de estudio propias de la escuela secundaria y las de la universidad, en particular este curso; (II) dificultades, particularmente en los primeros meses, para mantener el ritmo de estudio; (III) frecuentes impedimentos para avanzar en la resolución de los ejercicios debidas a la falta de conocimiento de los detalles del lenguaje de programación.

En [5] se analizó cómo los alumnos iban construyendo demostraciones formales, qué estrategias empleaban y qué recursos utilizaban; develando que "la construcción de una prueba no sigue el carácter lineal que posee una prueba una vez finalizada". Para los estudiantes elaborar una demostración era un proceso con idas y vueltas donde se recurría a compañeros y profesores y se utilizaban artefactos, principalmente el listado de axiomas del cálculo. El proceso implicaba realizar cálculos auxiliares, probar distintos caminos –algunos infructuosos– y la discusión con los colegas del curso.

Teniendo en cuenta estas dificultades, se inició el proyecto Theona [1] con el objetivo de desarrollar material pedagógico y herramientas didácticas informáticas que faciliten los procesos de aprendizaje de los alumnos en el curso. En estos dos años se desarrolló, financiados parcialmente por Fonsoft, el material de texto y cuatro herramientas didácticas:

Equ permite realizar demostraciones, automáticamente verificadas, de fórmulas ecuacionales.

Sat ayuda a comprender el significado de fórmulas lógicas de primer orden que involucren cuantificadores a través la adecuación de mundos de objetos geométricos y propiedades de dichos modelos expresados en fórmulas lógicas.

Fun permite derivar programas funcionales a partir de especificaciones formales; y también verificar la corrección de programas.

Hal es un asistente de construcción de programas imperativos con el uso de sistemas asercionales.

2.1. Introducción a la programación funcional

La curricula a abordar con la herramienta consta de dos unidades temáticas: introducción a elementos de lógica y especificaciones y derivación de programas

funcionales. El motivo de esta elección es que el proceso de desarrollo de software en la pequeña escala puede basarse en la especificación del problema en una fórmula lógica (complicada, por cierto) y que a partir de esa fórmula se pueden realizar manipulaciones simbólicas para obtener otra expresión formal: el programa que resuelve en forma algorítmica el problema especificado en la fórmula. Otra perspectiva que admite este curso es la verificación de programas: dada una especificación y un programa (digamos programado por otro estudiante), cómo podemos convencernos que el mismo satisface su especificación. Para ello, podemos utilizar la misma maquinaria lógica para demostrar que el programa efectivamente hace lo que la especificación prescribe.

Lógica para especificaciones La primera unidad temática está dedicada a los elementos de la lógica necesarios para poder escribir especificaciones. El énfasis está puesto en los elementos que componen un lenguaje formal, primero a través de la introducción de un lenguaje de expresiones aritméticas, para pasar luego a un lenguaje de fórmulas y el sistema de pruebas de esta lógica. Notemos que en este contexto el estudiante se enfrenta por primera vez a ciertos elementos que son comunes a cualquier lenguaje de programación: constantes, operadores y su aridad, variables, un elemental sistema de tipos y una gramática que definen las frases válidas del lenguaje.

Pensamos que esta primera aproximación a los fenómenos sintácticos en una lógica proposicional tiene la ventaja de no lidiar directamente con un lenguaje de programación. La elección del lenguaje de expresiones aritméticas tiene como ventaja que uno puede explicar informalmente la semántica de las expresiones (y de las fórmulas) sin necesidad de introducir nociones de modelos; la noción de pruebas es, básicamente, una secuencia de ecuaciones justificadas por proposiciones (ya sean axiomas, teoremas o hipótesis), donde cada paso ecuacional es correcto si los términos de la ecuación coinciden sintácticamente con la justificación. También en este momento se discuten las nociones de satisfacción y validez, nuevamente pensando en la semántica natural de las expresiones aritméticas.

En una segunda etapa se extiende la lógica proposicional a lógica de predicados tipada; ésta incluye cuantificadores lógicos y aritméticos, tales como la sumatoria, la productoria y el conteo. En este pasaje aparecen nuevos conceptos sintácticos que también son comunes a lenguajes de programación, en particular la noción de ocurrencias libres y ligadas de variables, renombre de variables ligadas, sustitución.

Con las expresiones cuantificadas se introducen las especificaciones de funciones que, a través de razonamientos ecuacionales, serán transformados en expresiones del lenguaje de programación funcional.

Programación funcional El lenguaje de programación, descrito en la sec. 3.1, permite la declaración de funciones a través de constantes, operadores, pattern-matching y recursión mutua. La noción de computación en este tipo de lenguajes se basa en la reducción de expresiones hasta llegar a los valores, o formas canónicas. Este paradigma de programación tiene la ventaja de ser cercano a la práctica matemática que se tiene habitualmente, en tanto el orden de las reducciones no afecta el valor final de la computación, puesto que no hay una manipulación

de un estado. Otra buena razón, a nuestro entender, para utilizar lenguajes de programación funcionales para introducir las nociones elementales es la ausencia de variables de estado, en este sentido las variables son realmente variables matemáticas cuyo valor no varía temporalmente.

3. Conceptos básicos

En esta sección daremos un descripción general de los conceptos a abordar mediante la utilización de la herramienta. Estos conceptos se fundamentan en lo descripto en la sección anterior.

3.1. Lenguaje de programación

El lenguaje de programación que utilizaremos es un lenguaje funcional que una sintaxis similar a Haskell. Este lenguaje incluye expresiones aritméticas y expresiones booleanas proposicionales. Por ejemplo, se puede definir la función de elevar al cuadrado de la siguiente manera

$$\texttt{cuadrado} : Int \to Int$$
$$\texttt{cuadrado}.x = x * x$$

La programación funcional se funda, sobre todo, en la definición de funciones sobre tipos inductivos; para lo cual se utilizan las definiciones recursivas. Los tipos inductivos que incluye el lenguaje son los naturales y las listas. El lenguaje, al igual que Haskell, permite definiciones utilizando patrones (pattern matching). Por ejemplo para sumar todos los elementos de una lista y teniendo en cuenta que las listas son tipos inductivos, definidos a partir de un caso base y un caso inductivo, la función `sum` puede ser definida utilizando esa estructura.

$$\texttt{sum} : [Int] \to Int$$
$$\texttt{sum}.[] = 0$$
$$\texttt{sum}.x \triangleright xs = x + sumar.xs$$

Existen una serie de operadores predefinidos para las expresiones aritméticas, booleanas y para operar sobre listas. Ejemplos de operadores sobre listas son concatenar dos listas ($+\!\!+$), calcular la cantidad de elementos de una lista ($\#$), etc.

A cada definición que hemos dado, la acompañan expresiones que determinan el tipo de la función. La utilización de un sistema con chequeo estático de tipos ayuda a detectar errores, pero a la vez ayuda a la legibilidad de la función. Por otro lado, desde el punto de vista didáctico, el tipado de expresiones y funciones es una herramienta importante con la que cuentan los estudiantes para la elaboración de nuevas funciones.

3.2. Sistema formal

El objetivo de un sistema formal es explicitar un lenguaje en el cual se realizarán demostraciones y las reglas para construirlas. Esto permite tener una noción muy precisa de lo que es una demostración, así también como la posibilidad de hablar con precisión de la sintaxis y la semántica de las expresiones y fórmulas del lenguaje. En una materia introductoria a la programación estos conceptos son importantes, dado que un lenguaje de programación es un sistema formal.

El diseño de este sistema formal permite, por un lado, demostrar la corrección de programas y derivar programas a partir de su especificación, ya que el lenguaje de programación está incluido en el mismo; y por otro lado, nos permitió desarrollar una herramienta automática de verificación de demostraciones y derivaciones de los programas.

Además de las expresiones aritméticas, booleanas proposicionales y sobre listas incluidas en el lenguaje de programación, se agregan las expresiones cuantificadas; que son de mucha utilidad para la especificación de programas y propiedades. Una expresión cuantificada tiene la forma $\langle \bigoplus x : R.x : T.x \rangle$ que se entiende como $T.x_0 \oplus T.x_1 \oplus T.x_2 \oplus \ldots$ donde x_i satisface el predicado R. Los cuantificadores que utilizados son el para todo ($\forall$), el existe ($\exists$), la sumatoria ($\sum$), la productoria ($\prod$) y el contador (N).

Para especificar, por ejemplo, la función `sum` que se definió recursivamente en la sección anterior, se puede describir de la siguiente manera

$$\texttt{sum.xs} = \langle \sum i \ : \ 0 \leqslant i < \#.xs \ : \ xs.i \rangle$$

3.3. Reglas de inferencia y demostraciones

Para completar el sistema formal se definen una serie de axiomas y teoremas básicos, demostrados a partir de los axiomas, y un sistema deductivo basado en la lógica ecuacional al estilo de Dijkstra, que utiliza la transitividad y la regla de Leibniz como reglas de inferencia, y se introduce el método inductivo cuando están implicados tipos inductivos.

Los axiomas sobre los que se trabaja son un conjunto de fórmulas para la lógica proposicional y para la lógica de cuantificadores. Para la aritmética se axiomatizan algunas propiedades sobre los números naturales, necesarias a veces para demostraciones inductivas sobre naturales, y otras propiedades se dejan a criterio del estudiante, es decir, se pueden utilizar como paso en una demostración, pero no se verifican formalmente.

Una *demostración* dentro del sistema formal consiste en probar la validez de una fórmula mediante una serie de pasos justificados con axiomas y teoremas ya demostrados. A continuación se muestra un ejemplo particular de la aritmética:

$$\begin{array}{ll} & (x > 0) \lor (x \leqslant 0) \\ \equiv & \{ \text{ Aritmética } \} \\ & (x > 0) \lor \neg(x > 0) \\ \equiv & \{ \text{ Tercero excluido } (P \lor \neg P \equiv True) \ \} \\ & True \end{array}$$

3.4. El proceso de construcción de programas

En la actualidad es ampliamente aceptado que el proceso de construcción de programas debe dividirse en al menos dos etapas: la etapa de **especificación** del problema y la etapa de desarrollo del programa o de **programación**.

El resultado de la primera etapa es una *especificación formal* del problema, la cual sigue siendo abstracta (poco detallada) pero está escrita formalmente. La segunda etapa da como resultado un programa y una demostración de que el programa es *correcto respecto de la especificación dada*. A esta demostración se la llama verifiación.

Por ejemplo, a continuación se muestra la demostración de la corrección de la función `sum` cuya especificación y definición recursiva se introdujo anteriormente.

La demostración consiste en una prueba que, para cualquier lista xs, la función `sum` satisface la especificación. Esta demostración se puede hacer a partir de los axiomas y teoremas básicos del formalismo y utilizando el principio de inducción. El *caso base* supone que la lista es vacía. En esta situación la especificación se convierte en $\sum.[\,] = \langle \sum i : 0 \leqslant i < \#[\,] : [\,]!i \rangle$ que se demuestra trivialmente. Para el *caso inductivo* la especificación toma la forma:

$$sum.(x \triangleright xs) = \langle \sum i : 0 \leqslant i < \#(x \triangleright xs) : (x \triangleright xs)!i \rangle$$

que se demuestra a continuación

$$\begin{array}{ll} & \langle \sum i : 0 \leqslant i < \#(x \triangleright xs) : (x \triangleright xs)!i \rangle \\ = & \{ \text{ Definición de } \# \ \} \\ & \langle \sum i : 0 \leqslant i < 1 + \#xs) : (x \triangleright xs)!i \rangle \\ = & \{ \text{ Teorema separación del primer término } \} \\ & (x \triangleright xs)!0 + \langle \sum i : 0 \leqslant i < \#xs : (x \triangleright xs)!(i+1) \rangle \\ = & \{ \text{ Definición de } ! \ \} \\ & x + \langle \sum i : 0 \leqslant i < \#xs : xs!i \rangle \\ = & \{ \text{ Hipótesis inductiva } \} \\ & x + \texttt{sum}.xs \\ = & \{ \text{ Definición de } \texttt{sum} \ \} \\ & \texttt{sum}.(x \triangleright xs) \end{array}$$

3.5. Derivación de programas

Hasta aquí se desarrollaron tres etapas para la construcción de un programa. En primer lugar, elaborar una especificación formal para el mismo. En segundo lugar, construir el programa. En tercer lugar, dar una demostración de que dicho programa satisface la especificación.

A la construcción en simultáneo del programa y de su verificación se la llama *derivación*. En este proceso se parte de una especificación formal de una función y a través de transformaciones de la expresión se encuentra la definición de la función. Esta metodología fue desarrollada a partir de los trabajos de E. W. Dijkstra.

En la sección 4 se describe un ejemplo de derivación utilizando la herramienta.

4. FUN : una herramienta didáctica

FUN es una herramienta que consiste en un IDE (entorno de desarrollo integrado) para escribir en un lenguaje de programación y de demostraciones matemáticas. Este lenguaje permite definir progamas funcionales, especificaciones de programas, realizar pruebas matemáticas utilizando lógica proposicional y de predicados, y realizar derivaciones de programas de acuerdo con los conceptos introducidos en la sección anterior.

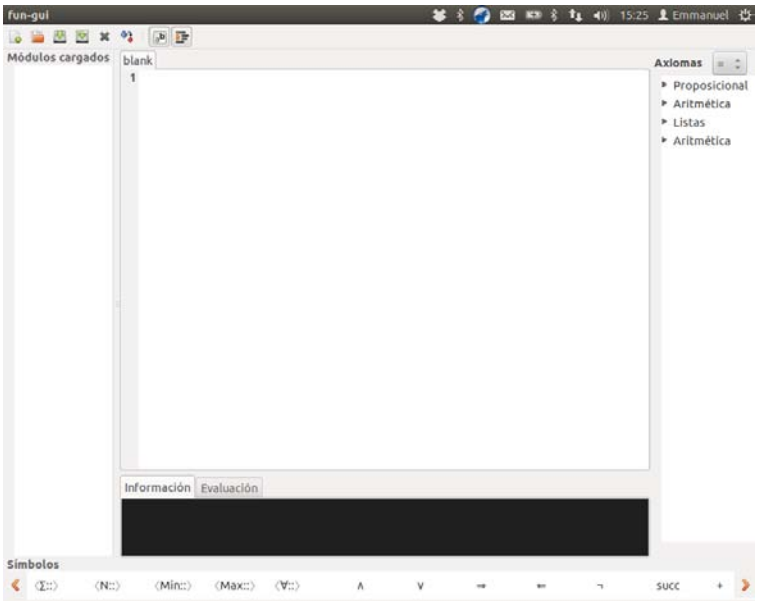


Figura 1: La pantalla principal de FUN

En la sección central de la pantalla se muestra un editor de textos, a la derecha está la lista de axiomas disponibles para utilizar como justificación en las pruebas, en la sección de abajo se ve una consola de información y otra de evaluación, y más abajo botones para poder ingresar caracteres unicode fácilmente.

El programa permite escribir y chequear pruebas del cálculo proposicional y de predicados. Las pruebas se realizan mediante la declaración de **teoremas**. Por ejemplo consideremos el teorema de doble negación $\neg\neg p \equiv p$, ver la Fig. 2b. Para realizar la prueba de $\neg\neg p \equiv p$ se utiliza un teorema auxiliar `teo1`, Fig. 2a. Para justificar los pasos de la prueba, el usuario puede utilizar una cantidad de axiomas seleccionándolos mediante la interfaz o escribiendo literalmente sus nombres, puede utilizar teoremas ya probados previamente o hipótesis, y también puede utilizar definiciones de funciones.

Una vez que se escribe un módulo (un archivo que contiene definiciones de funciones, demostraciones y/o especificaciones) el usuario puede chequear que el

```
1 module TeoremasEjemplo
2
3 let thm teo1 =
4     ¬ False ≡ True
5
6 begin proof
7   ¬ False
8   ≡ { Neutro de la equivalencia a derecha }
9   ¬ False ≡ True
10  ≡ { Negación y Equivalencia }
11  ¬ (False ≡ True)
12  ≡ { Conmutatividad de la Equivalencia }
13  ¬ (True ≡ False)
14  ≡ { Negación y Equivalencia }
15  ¬ True ≡ False
16  ≡ { Conmutatividad de la Equivalencia }
17  False ≡ ¬ True
18  ≡ { Definición de False }
19  True
20 end proof
```

(a) Teorema auxiliar

```
23 let thm dobleNeg =
24   ¬ (¬ p) ≡ p
25
26 begin proof
27   ¬ (¬ p)
28   ≡ { Neutro de la equivalencia a izquierda }
29   ¬ (¬ (True ≡ p))
30   ≡ { Negación y Equivalencia }
31   ¬ (¬ True ≡ p)
32   ≡ { Definición de False }
33   ¬ (False ≡ p)
34   ≡ { Negación y Equivalencia }
35   ¬ False ≡ p
36   ≡ {|teo1 }
37   True ≡ p
38   ≡ { Neutro de la equivalencia a izquierda }
39   p
40 end proof
```

(b) Doble negación

Figura 2: Demostraciones de teoremas

mismo sea correcto y la herramienta mostrará la lista de todas las declaraciones indicando con una tilde si la misma es correcta, Fig. 3a o con una cruz cuando no lo es, Fig. 3b. Si alguna declaración tiene un error, cuando se hace click sobre la misma en el panel izquierdo, en la consola de información se muestra un mensaje de la razón por la cual la misma no es correcta, Fig. 3c.

La derivación de funciones en el entorno FUN sigue el mismo proceso que se explicó en la sección anterior; en las Figs. 4a y 4b se muestran respectivamente los casos bases e inductivos de la derivación de la sumatoria de una lista. Como se puede ver, en ambos casos la expresión final no contiene cuantificadores y por lo tanto es un programa evaluable.

El lenguaje funcional subyacente a FUN fue definido y estudiado formalmente en [4]; utilizando la semántica operacional allí definida se implementó el evaluador integrado en el entorno. El usuario puede evaluar una expresión paso-a-paso hasta la expresión canónica.

5. Conclusiones y trabajos futuros

En el artículo hemos presentado una metodología para la enseñanza de programación enfatizando la importancia de contar con una especificación formal. El aporte más novedoso de este artículo es la herramienta FUN, que permite al estudiante realizar la derivación, programación y verificación en un único entorno. La herramienta le brinda al estudiante la posibilidad de corregir sus propias derivaciones sin necesidad de la supervisión de un docente.

En el próximo año se espera utilizar el material de estudio producido y las herramientas informáticas en el dictado del curso “Introducción a los algoritmos”.

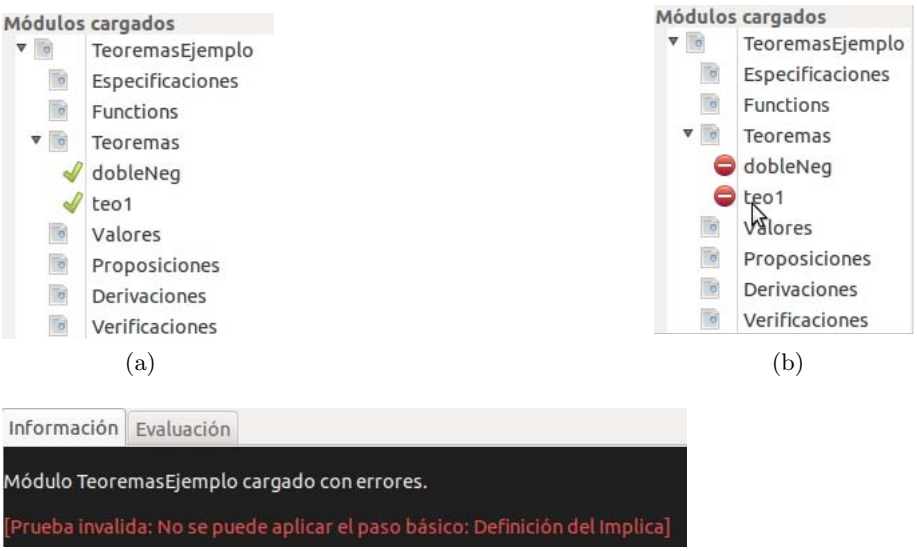


Figura 3: Módulos y sus definiciones

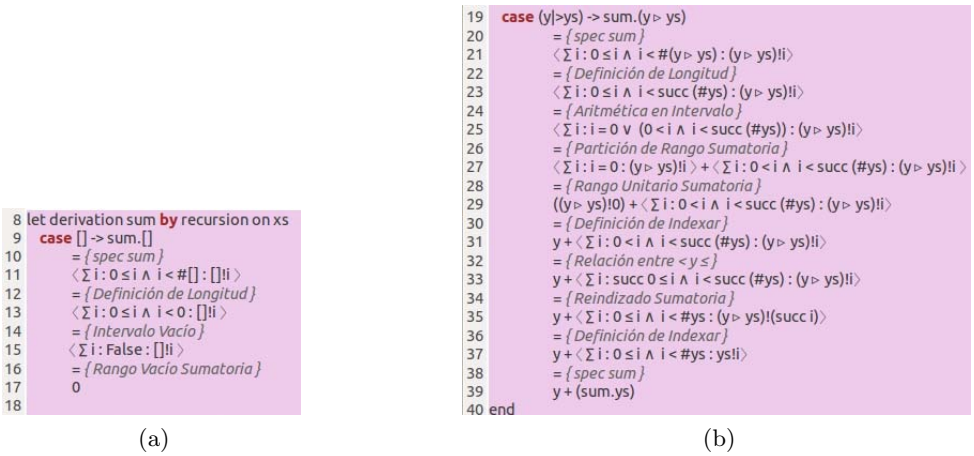


Figura 4: Derivación de `sum`

A partir del uso intensivo de las herramientas se contará con un importante feedback para mejorarlas, tanto en funcionalidades como en interfaz.

En otras líneas de trabajo se están desarrollando herramientas que complementan estos contenidos curriculares. En particular, se está desarrollando una herramienta, Hal, que persigue los mismos objetivos que FUN pero orientada a la programación imperativa; esto implica un cambio completo de paradigma. Esta herramienta permitirá completar con los contenidos conceptuales de un curso introductorio de programación desde la perspectiva formal.

Referencias

[1] Araceli Acosta y col. *Proyecto Theona*. http://www.theona.com.ar/. Mayo de 2013.

[2] Javier Blanco, Silvina Smith y Damián Barsotti. *Cálculo de Programas*. Universidad Nacional de Córdoba, 2009. ISBN: 978-950-33-0642-0.

[3] Javier Blanco y col. "An introductory course on programming based on formal specification and program calculation". En: *SIGCSE Bull.* 41.2 (jun. de 2009), págs. 31-37. ISSN: 0097-8418. DOI: 10.1145/1595453.1595459.

[4] Emmanuel Gunther. "Entorno para la Derivacion de Programas". Tesis de lic. Universidad Nacional de Córdoba - Facultad de Matemática, Astronomía y Física, jul. de 2013.

[5] Leticia Losano. "Procesos situados de aprendizaje en cursos básicos deprogramación: volverse miembro de una comunidad". Tesis doct. Universidad Nacional de Córdoba - Facultad de Filosofía y Humanidades, abr. de 2012.

[6] Mariëlle Stoelinga y Ralf Pinger, eds. *Formal Methods for Industrial Critical Systems*. Vol. 7437. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-32468-0. DOI: 10.1007/978-3-642-32469-7.

Metodología innovadora para el Estudio y Programación de Microprocesadores en Arquitectura de Computadoras.

Jorge R Osio, Daniel Alonso, Eduardo Kunysz, Martín Morales

Universidad Nacional Arturo Jauretche

Instituto de Ingeniería – Ingeniería Informática
Av. Clachaquí 6200, Florencio Varela, Argentina
josio@unaj.edu.ar, martinmorales@unaj.edu.ar

Abstract. Este Trabajo presenta una Metodología de enseñanza de Microprocesadores, con una etapa de programación en Assembler, la cual facilita la comprensión en el funcionamiento interno de un microprocesador y el rol que tienen los registros en el mismo. Adicionalmente, se plantea una etapa de aplicación de los Microprocesadores en el Diseño de Sistemas Digitales mediante Programación de software para sistemas embebidos. Dicha metodología se está implementando en la Asignatura de Organización y Arquitectura de Computadoras.
La propuesta se encuadra dentro del Método inductivo básico a través de Simulación con kits e instrumental de diseño de Sistemas Embebidos. Este método se considera apropiado, teniendo en cuenta el contenido a enseñar y la necesidad de incluir la enseñanza de Programación de Sistemas Embebidos basado en sistemas Procesadores en las carreras de Informática. La Programación de Software embebido en sistemas Microprocesadores, se presenta como un Trabajo de Laboratorio que se va desarrollando a lo largo de toda la materia.

Keywords: Organización y Arquitectura de Computadoras, Microprocesadores, Sistemas Embebidos, Metodologías innovadoras, Estrategias Educativas.

1 Introducción

Para una mejor descripción de la metodología de enseñanza, se comienza por realizar una descripción del área dentro de la especialidad en donde se encuentran incluida la materia. Luego de esto, se describe la metodología en sí misma, acompañada de los contenidos de la materia.

El área de enseñanza se denomina ARSO y abarca las materias de Arquitectura, Redes y Sistemas Operativos, las cuales están fuertemente vinculadas al Hardware dentro de las Ciencias Informáticas.

Entre los Objetivos y Contenidos se busca describir la importancia de introducir al alumno en el funcionamiento y constitución del Microprocesador y mediante la programación del mismo a bajo nivel, acompañando los fundamentos teóricos con una fuerte parte práctica que permite a los alumnos desarrollar y ejercitar la capacidad de diseño de un programa y la resolución de problemas. Por otro lado se pretende introducir el concepto de software embebido mediante la realización de un Laboratorio a lo largo de toda la cursada, mediante un kit de desarrollo basado en un sistema microprocesador Cortex M3, que permite potenciar la programación de microprocesadores en alto nivel para aplicaciones complejas que involucran la implementación de Sistemas Operativos de Tiempo Real (RTOS) y Aplicaciones con interfaces Ethernet, para potenciar los conceptos adquiridos en redes.

La metodología presentada pretende demostrar cuan eficiente es aplicándola en el área ARSO mediante el respaldo de un grupo docente fuertemente capacitado para dar soporte y ayudar al alumno a desarrollar la capacidad de razonamiento sobre los temas en cuestión. Dicha Metodología se está aplicando actualmente en la material Organización y Arquitectura de Computadores del Instituto de Ingeniería de la UNAJ.

2 Descripción del Área

Las materias básicas del Área, son “Organización y Arquitectura de Computadoras” y “Sistemas Operativos I” dictadas en el tercer semestre de la carrera, luego “Redes de Computadora I” se dicta en el cuarto semestre, el resto de las materias del área se ubican en tercer y cuarto año de Ing. Informática. La aplicación de la metodología se centra en Organización y Arquitectura de Computadoras, en donde se contempla la descripción de un microprocesador y sus aplicaciones, además se presentan conceptos de programación de un sistema procesador en lenguaje de alto nivel con

aplicaciones relacionadas a las demás materias del área, todo esto mediante un Laboratorio obligatorio a realizarse durante toda la materia. Cabe aclarar que con las aplicaciones no se pretende explicar temas que serán impartidos en otras Materias, pero se abre una puerta para profundizar sobre estos temas de manera práctica y aplicada.

3 Objetivos y Contenidos

Los objetivos en relación a los contenidos de la materia "Organización y Arquitectura de Computadoras" son los siguientes:

- **Matemática Discreta**
  - Sistemas de Numeración
  - Algebra de Boole
  - Sistemas lógico y compuertas

- **Comprensión del funcionamiento de un procesador.**

Estudio desde el punto de vista físico y lógico de los microprocesadores:

  - Diagrama de bloques. Buses. Registros. Instrucciones. Modos de direccionamiento. Estructura algorítmica. (CPU)
  - Periféricos de entrada/salida. Proceso de interrupción. Temporizadores. Comparadores y capturadotes.(MCU)

- **Tipos y Selección de procesadores genéricos**
  - Estado del arte y criterios comparativos de procesadores

- **Utilización de procesadores genéricos:**
  - Programación en Assembler mediante un Microrpocesador básico
    - Caso de Estudio HC08 de Freescale (facilita la comprensión de funcionamiento)
    - Aplicaciones de algoritmos matemáticos y planteo con Diagrama de Flujos
  - Programación en C, Realización de experiencias concretas con elementos de entrada/salida
    - Caso de estudio Microprocesador Cortex M3
    - Aplicaciones con periféricos.

- **Descripción de las Distintas Arquitecturas de Computadores**
  - Evolución y Performance de las Computadoras

- **Memorias**
  - Memoria Cache
  - Memoria Interna
  - Memoria Externa

- **Entradas / Salidas**

- **Soporte para Sistemas Operativos**

- **Estudio de Arquitecturas de Microprocesadores populares**
  - Arquitecturas X86
  - Arquitectura ARM para Sistemas Embebidos

4 Materiales de Estudio y Herramientas de trabajo

Para la implementación de esta metodología es estrictamente necesario lograr una estrecha relación entre los materiales de estudio y las herramientas de trabajo. Es por eso que en cada Trabajo práctico siempre se agrega la información necesaria para que el mismo pueda ser resuelto sin problemas. Por otro lado el hecho de usar herramientas específicas de HW, hace necesario proveer materiales específicos para cada herramienta, ya sea, mediante apuntes de cátedra o mediante manuales o notas de aplicación relacionadas con las herramientas.

4.1 Materiales de Estudio

- Bibliografía básica [1 - 8]
- Apuntes de cátedra [9]
- Manuales Técnicos y notas de aplicación del microcontrolador HC908 y del kit de Desarrollo LPC1769 [10 - 14]
- Bibliografía de aplicaciones [15] y [16]

4.2 Herramientas de Trabajo

Dentro de las herramientas de trabajo se deben incluir la Computadora, como una herramienta indispensable para realizar toda la parte práctica.

4.2.1. Herramientas de Trabajo

Las herramientas de trabajo indispensables para llevar a cabo la metodología son:

- Kit de Desarrollo LPC1768, (ver fig. 1)
- Placa Base con los siguientes componentes:
    - Componentes para implementación de diferentes protocolos (serial, SPI, I^2C,PWM, USB, Ethernet)
    - Componentes básicos (leds, potenciómetros, dimmers, joystick)
    - Micrófono y parlantes para aplicaciones de audio
    - Memoria Flash externa
    - Interfaz con display gráfico
    - Interfaz USB (HOST y HID)
    - Interfaz Ethernet
    - Puertos de E/S

- Software para diagramas de Flujo.
- Software de compilación y simulación en circuito (Win IDE, ver Fig. 2)
- Software IDE LPCxpresso, Permite crear proyectos en C para Microcontroladores de 32 Bits basado en Microprocesadores ARM de la Firma NXP, (ver Fig. 3). Integra un conjuntos de Herramientas que permiten compilar y hacer debugging en circuito, también permite crear proyectos a compilar con el software Mingw.

5 Características de la Cursada

5.1 Contenidos Teóricos de la Cursada

En Principio se describen los componentes de una Computadora y se presenta la evolución y desempeño de las mismas.
Se detallan cada una de las partes Principales que la conforman como las distintas memorias (memoria cache, Memoria interna y Memoria externa) y se describen las entradas y salidas. También se explica el soporte para Sistemas operativos.
También, se dictan los detalles del Procesamiento, Como la Aritmética Computacional, Características del Set de Instrucciones, Modos de direccionamiento y la Unidad de Control. Esto último tanto para la Arquitectura HC08 de Freescale, luego se comparan características de X86 de Intel y de la Arquitectura ARM.
Luego se describen los Registros, Modos de Direccionamiento, Bifurcaciones, Saltos y subrutinas e Interrupciones del Microprocesador HC08 de Freescale con Registros que manejan datos de 8 bits, para la programación en Assembler. En una segunda Etapa se describen los Módulo de Entrada/Salida, Módulo de temporización, Comunicaciones Serie (SPI, SCI, I2C, Ethernet, USB, entre otros).

5.1 Contenidos Prácticos de la Cursada

- TP1. Rendimiento de Computador
- TP2. Sistemas de Numeración
- TP3. Algebra de boole
- TP4. Circuitos Lógicos y Partes de Microprocesador. Concepto de Procesamiento de Instrucción

- TP5: Ejercicios sobre Modos de direccionamiento y set de instrucciones basados en el HC08.
- TP6: Programación de operaciones basadas en los principales sistemas de numeración y algoritmos matemáticos con simulación en Assembler. (Realización de Diagramas de Flujo) sobre el HC08.
- TP7: Memoria externa. Discos Rígidos. RAID. I/O. Manejo de Memoria y paginado
- Laboratorio Obligatorio de Programación de Interfaces de e/s sobre una Arquitectura ARM en C. Este laboratorio consiste en la explicación de las interfaces que se enumeran a continuación y culmina con una aplicación que involucra las distintas interfaces sobre la placa LPC 1769.
  - En este Laboratorio se utiliza el kit de programación junto con una Placa Base, permitiendo implementar una comunicación RS-232, USB y Ethernet.
  - Aplicaciones con Display Gráfico
  - Realización el control de un sistema calefactor – refrigerante mediante el sensado de temperatura y el accionamiento de una resistencia calefactora o un ventilador según el valor de temperatura sensado.
  - Realizar aplicaciones de almacenamiento de datos mediante una memoria externa mediante la interfaz SPI
  - Realizar aplicaciones de procesamiento digital de audio.

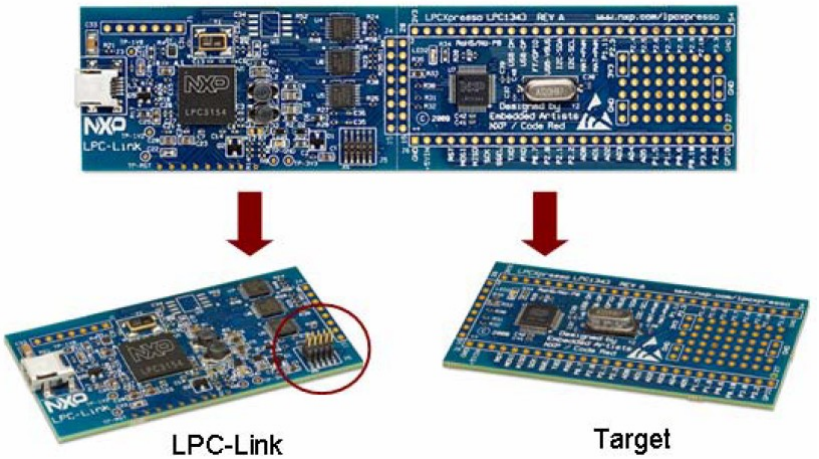


Fig. 1. El KIT de desarrollo LPCxpresso LPC1114 está formado por el LPC-Link, el cual es un programador que permite programar toda la Familia de Microcontroladores de NXP y hacer debugging en Circuito. Adicionalmente viene acompañado por un Target que contiene un Microcontrolador LPC1114 con un Microprocesador ARM Cortex M0 para Aplicaciones.

6 Metodología de Enseñanza

La metodología de enseñanza tiene en cuenta cuatro factores principales para la instrucción del Alumno:

- El primer factor consiste en hilvanar los temas vistos en materias anteriores. Esto se logra, integrando los conocimientos vistos anteriormente en un Trabajo Práctico, mediante ejercicios que involucran operaciones y algoritmos matemáticos.
- De la misma manera se integran los Contenidos de "Organización y Arquitectura de Computadores" con "Sistemas Operativos I", Aprovechando la implementación de un Sistema Operativo de Tiempo real en el Laboratorio, aplicando en la Práctica los conceptos adquiridos en dicha materia.
- Los dos Factores principales están muy estrechamente ligados y coordinados para lograr instruir al alumno de manera que pueda fijar los conocimientos teóricos mediante la aplicación práctica de los mismos. Por un lado, la programación en Assembler de un microprocesador básico fortalece los conceptos de funcionamiento de un procesador y las posibles aplicaciones de sus registros. Por otro lado, involucra la aplicación de las reglas del buen programador mediante la planificación del código mediante el Diagrama de Flujos correspondiente.
- Por último, se puede considerar como factor complementario el hecho de incluir dentro del Laboratorio, aplicaciones que involucran temas y elementos específicos de otras materias de Informática, mediante la Programación en Lenguaje de Alto nivel de código para Sistemas embebidos, la realización de aplicaciones que contienen elementos de materias del área ARSO, etc.

6.1 Metodología de Teoría

Las clases teóricas se dictan mediante filminas ilustrativas que permiten al alumno apreciar detalles de Imágenes y gráficos de manera interactiva. Los temas teóricos están diagramados de tal manera que los alumnos puedan fijarlos mediante la correspondiente clase práctica, de manera casi simultánea.

6.2 Metodología Práctica

La metodología práctica está diagramada en dos módulos, como se detalla en la sección 5. En el primer Módulo se presenta una práctica que consiste en determinar mediante diferentes técnicas la eficiencia y rendimiento en sistemas Computadores. Introducir al alumno en los sistemas de numeración, especialmente los utilizados en sistemas digitales. Enseñar al alumno a representar las distintas partes de un microprocesador mediante compuertas lógicas y por último, introducción al alumno al funcionamiento de un Computador, mediante la descripción de la composición y funcionamiento de un Microprocesador, de las jerarquías de memoria y de los periféricos I/O.
Los Trabajos Prácticos 5, 6 y el laboratorio se realizan en Computadora con diferentes herramientas de software. En cada clase los alumnos tienen una introducción a las prácticas dictada por el Profesor a cargo del Curso y luego los alumnos realizan parte de los ejercicios correspondientes al Trabajo práctico en cuestión. La primera parte del segundo módulo consiste en programar código assembler profundizando sobre las características y funcionamiento de un microprocesador (en cuanto a registros, set de instrucciones, códigos de operación y modos de direccionamiento). En esta instancia el alumno comienza a utilizar el software WINIDE, de la Fig. 2, mediante la compilación y la simulación de código. Luego sigue una Práctica que consiste en la programación y simulación en assembler de ejercicios que requieren el manejo de datos, la conversión entre sistemas de numeración, implementación de operaciones matemáticas y de algoritmos matemáticos y de comunicaciones. Cabe aclarar que previo a cada programa se deben realizar el diagrama de flujo describiendo el funcionamiento que deberá realizar el programa. Este tipo de ejercicios hace que el alumno desarrolle la capacidad de programar de manera ordenada y de resolver problemas.

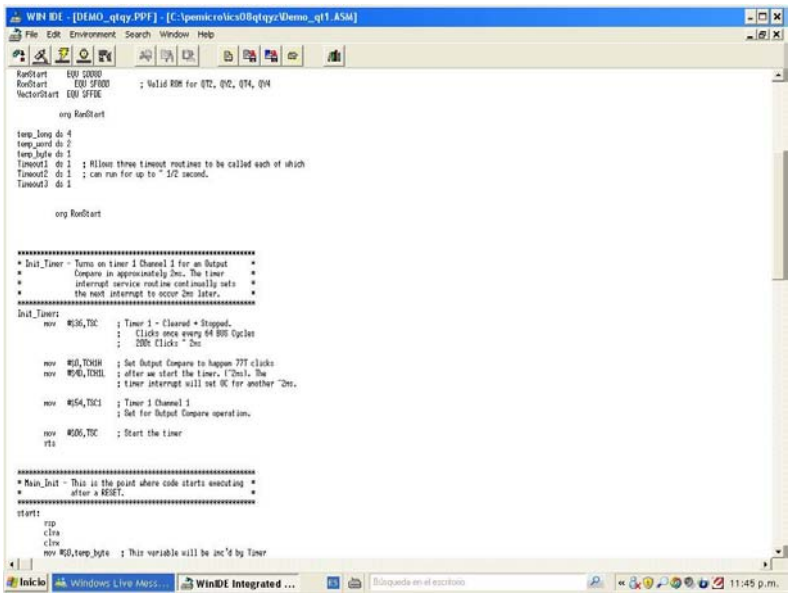


Fig. 2. Compilador y Simulador en Circuito WINIDE.

Es importante destacar que los alumnos forman comisiones de tres integrantes para realizar El Laboratorio, lo cual fortalece el trabajo en grupo, en donde además de la aplicación a implementar, deben realizar un informe y presentarlo en una clase al resto del curso.
El Laboratorio, consiste básicamente en la programación en Lenguaje C de los módulos periféricos E/S que permiten la interfaz entre el microprocesador y otros dispositivos. De esta manera se realiza el control de un Display gráfico, de varios módulos externos como memorias flash mediante interfaz I^2C, interfaz USB, interfaz Ethernet, Free RTOS y procesamento de audio mediante PWM, etc. Para la programación del Microcoprocesador Cortex M3 se utiliza el software LPC xpresso que se muestra en la Fig. 3.

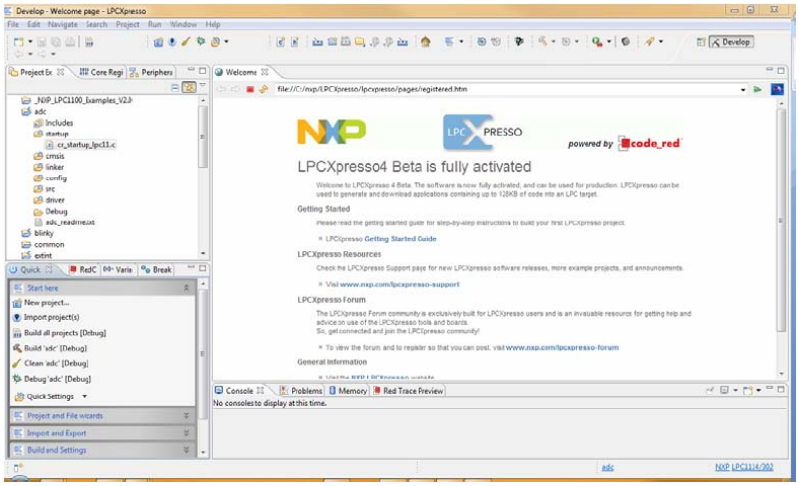


Fig. 3. Entorno de Desarrollo Integrado LPCxpresso, que permite Programación en C y debugging en circuito para Microcontroladores de NXP basados en Microprocesadores ARM Cortex.

Finalmente, con el Laboratorio se introduce al Alumno a la programación de Software para sistemas Embebidos, se presentan aplicaciones relacionadas con Sistemas operativos de tiempo real y Ethernet, permitiendo aplicar los conceptos profundizados en las demás materias del área. Se implementa el envío de datos a una memoria externa, aplicando los conceptos de jerarquía de memoria. Por último, se realizan aplicaciones con periféricos I/O, lo cual permite comprender la idea de interfaz E/S en un Computador y buses y se exploran las distintas posibilidades que ofrece un Sistema procesador ARM.
Se debe aclarar que cada curso se compone de un profesor a cargo y un máximo de 30 alumnos, es por eso que se puede llevar a cabo esta modalidad tan demandante.

7 Resultados Obtenidos

Entre los resultados obtenidos se puede destacar que la programación en Assembler de un microprocesador de 8 bits con registros básicos favorece la comprensión del funcionamiento del mismo y las posibles funciones que pueden cumplir los registros. De un total de 120 alumnos repartidos en 4 cursos de 30, se presentaron a los exámenes 104 alumnos, sobre los cuales se obtuvieron los resultados de la Tabla 1, en donde se muestran los resultados luego de las evaluaciones correspondientes a los temas relacionados con el microprocesador, tales como modos de direccionamiento, registros, set de instrucciones y ejecución de instrucciones, programación en Assembler.

Tabla 1. Resultados de las evaluaciones del Segundo módulo referentes al microprocesador y su funcionamiento.

Tema evaluado	Sin Respuesta	Respuesta correcta	Respuesta incorrecta
Descripción teórica de ejecución de instrucciones en un microprocesador	4%	90%	6%
Realización de diagramas de flujo	40%	45%	15%
Programación en Assembler	15%	60%	25%
Modos de direccionamientos y set de instrucciones	2%	90%	8%
Utilización de registros y memoria	15%	65%	20%

En la Tabla 2 se puede observar como influyeron los conceptos aplicados mediante la programación de un microprocesador en Assembler, en los alumnos que tuvieron que rendir en la última fecha flotante de la materia el módulo 1.

Tabla 2. Comparativa de los temas relacionados con el microprocesador evaluados en el primer módulo antes y después de programar en Assembler y explorar las características del microprocesador .

Evaluación de conceptos sobre microprocesadores en el primer módulo	Sin Respuesta	Respuesta correcta	Respuesta incorrecta
Examen del primer módulo sobre un total de 115 alumnos	10%	60%	30%
Examen flotante del primer módulo sobre un total de 20 alumnos	5%	95%	0%

8 Conclusiones

Como conclusión se puede decir que esta metodología ayuda al alumno a familiarizarse con todas las herramientas digitales disponibles en el diseño y Programación de Sistemas Embebidos y le provee una buena preparación para desempeñarse como profesional en esta área.
Entre las Herramientas presentadas y utilizadas en la Cátedra se presentan las diferentes Familias de Microprocesadores que son usados en una amplia gama de aplicaciones para Sistemas Embebidos y Sistemas Computacionales.
Por un lado se realizan aplicaciones sobre un microcontrolador de 8 bits de baja gama como es el HC08, en donde se pueden visualizar y utilizar los principales registros de un procesador, mediante la programación en bajo nivel de aplicaciones que requieren poco procesamiento y poca memoria. Entre los resultados, se pudo demostrar que esta metodología favorece a la comprensión del funcionamiento del microprocesador y de cada una de sus partes. También se logró fomentar la planificación de un programa mediante diagramas de flujo y la resolución de problemas.
Aunque el Laboratorio de software embebidos sobre el kit LPCxpresso aun no nos permite sacar conclusiones por su reciente implementación, es muy evidente que fortalecerá la programación de sistemas embebidos y por otro lado formalizar los conocimientos sobre jerarquías de memorias y Periféricos I/O.

Referencias.

1. Cazares Juan, Haro Diego, Hueso Jaime, Muriel Eduardo, Puebla Luis: Microcontroladores Motorola - Freescale Programación Familias y Sus Distintas Aplicaciones en La Industria, Ed. Alfaomega 2008.
2. Programación de Sistemas Embebidos en C, Gustavo Galeano, Ed Alfaomega, 2009
3. Spasov Ed. Prentice Hall (5th Edition) 2004
4. D. A. Patterson, J. L. Hennessy: Embedded Microcomputer Systems Real Time Interfacing. Ed Thomson 2da Edition 2002
5. D. A. Patterson, J. L. Hennessy: Estructura y diseño de computadores. Ed. Reverté, 2000
6. Computer Organization and Architecture Designing for Performance (8th Edition) –William Stallings
7. Computer Organization and Architecture, (9th Edition) - William Stallings
8. Organización Diseño de Computadores. La interfaz hardware/software, Patterson, David A. – Henessy, John L., Mc Graw-Hill, 1995
9. Jorge R. Osio, Daniel Alonso, Eduardo Kunysz ” Decripción de un Microprocesador– CPU”, Instituto de Ingeniería, UNAJ, 2013
10. Data sheet: MC68HC908QY4 Microcontrolers, Rev. 5, 07/2005
11. Reference Manual: CPU Central Processor Unit Microprocesador, CPU08RM, Rev. 4, 02/2006.
12. The definitive guide to the ARM-Cortex M3, Joseph Yiu, Elsevier Inc, 2010
13. Getting Started with NXP LPCXpresso, Revisión 11.1, Diciembre de 2011
14. LPC111x/LPC11Cxx User manual, revisión 6, Agosto 2011
15. Douglas H. Summerville: Embedded Systems Interfacing for Engineers using the Freescale HCS08 Microcontroller II: Digital and Analog Hardware Interfacing‖, State University of New York at Binghamton, Morgan y Claypool Publishers, 2009.
16. Mark Martinets: Interrupt Handling Considerations When Modifying EEPROM on HC08 Microcontrollers‖, Nota de aplicación, Motorola, agosto de 2002.

Usando NDT como soporte a la enseñanza de programación web

Yanina Medina, Gabriel Pedrozo Petrazzini, Cristina Greiner, Gladys Dapozo
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura.
Universidad Nacional del Nordeste. Corrientes. Argentina
{gndapozo, cgreiner, yanina}@exa.unne.edu.ar, gabriel.pedrozopetrazzini@gmail.com

Abstract. Se presentan los resultados de la implementación de una metodología de enseñanza de programación para la plataforma web que utiliza la metodología NDT, llevada a cabo en una asignatura de tercer año de la carrera Licenciatura en Sistemas de Información de la Universidad Nacional del Nordeste (UNNE). Esta estrategia surge con el objetivo de afianzar en los alumnos el valor de las buenas prácticas que exige la Ingeniería del Software para lograr desarrollos y soluciones cada vez más completas y robustas. El proceso de aseguramiento de calidad tiene como misión principal garantizar todos los requisitos de calidad establecidos. Para ello, los controles de calidad no deben aplicarse únicamente al código generado, sino que además deben recorrer elementos como los requerimientos, tanto funcionales como no funcionales, que contribuyen a generar conciencia de la importancia que tiene la documentación en el desarrollo de software.

Keywords: Enseñanza universitaria**,** programación web, NDT, documentación.

1 Introducción

A pesar de los esfuerzos orientados a la creación de metodologías de desarrollo para la web, el uso sistemático de estas técnicas para la especificación y el diseño de estas aplicaciones no ha resuelto el problema de la producción. Por este motivo, los expertos en tecnologías web han realizado diferentes propuestas para mejorar la calidad de los sitios y aplicaciones web, en forma de metodologías, marcos de calidad, modelos de estimación, guías de estilos y métricas [1].

Las metodologías de desarrollo de software son un conjunto de procedimientos, técnicas, herramientas y un soporte documental que ayuda a los desarrolladores a realizar un producto software. Un caso particular, lo constituyen las metodologías orientadas al desarrollo web. Por sus características, estas requieren una mayor atención en la definición de los requerimientos funcionales y no funcionales, y dentro de estos últimos, a los requerimientos de almacenamiento y de navegabilidad.

En un estudio previo [2], se realizó una comparación de metodologías web, analizando en particular el grado de cobertura de las distintas etapas de desarrollo. De las metodologías estudiadas, únicamente NDT cuenta con soporte para todas las etapas del ciclo de vida.

Por este motivo, se eligió esta metodología de desarrollo de aplicaciones web, para ser utilizada en la enseñanza de la programación de aplicaciones web, con el objetivo de afianzar en los alumnos el valor de las buenas prácticas que exige la Ingeniería del Software para lograr desarrollos y soluciones cada vez más completas y robustas.

2 NDT

NDT (*Navigational Development Techniques*) es una metodología para especificar, analizar y diseñar el aspecto de la navegación en aplicaciones web [3][4]. El flujo de especificación de requerimientos de NDT comienza con la fase de captura de requerimientos y estudio del entorno, y luego se definen los objetivos del sistema. En base a estos objetivos, el proceso continúa definiendo los requerimientos que el sistema debe cumplir para cubrir los objetivos marcados. Finalmente, se realiza la revisión del catálogo de requerimientos y el desarrollo de una matriz de trazabilidad que permite evaluar si todos los objetivos han sido cubiertos en la especificación. En la Fig. 1 se muestra una descripción general de las actividades de NDT.

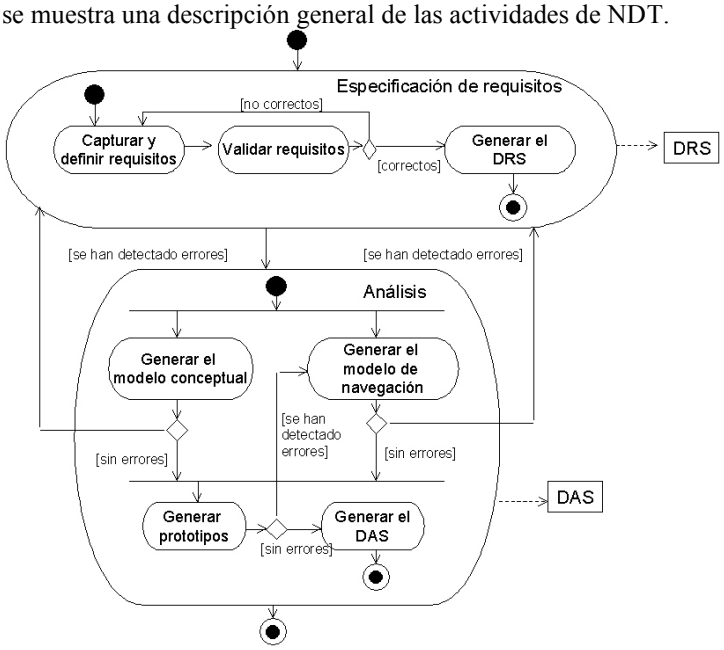


Figura 1. Descripción general de las actividades de NDT

2.1 Modelo de Requisitos

NDT es un enfoque específicamente creado para el manejo de requisitos de aplicaciones web [6]. Propone el uso de diagramas de casos de usos y varios tipos de plantillas de formato (patrones). Clasifica los requisitos en los siguientes tipos: de información de almacenamiento, de actores, funcionales, de interacción y requisitos no funcionales. Para cada tipo, se define una plantilla especial, consistente en una tabla con campos de texto específicos que son completados por el equipo de desarrollo en la fase de elicitación de requisitos.

2.2 Descripción de la herramienta: NDT-SUITE

NDT-Tools, el soporte de herramientas de la metodología NDT, ha tenido que evolucionar para ser una propuesta útil en proyectos reales, dado que sólo cubría las fases de ingeniería de requisitos y análisis [7] [8]. Estas razones impulsaron al Grupo de Investigación Ingeniería Web y Testing Temprano [5] a elaborar NDT-Suite. Esta nueva herramienta soporta las fases de requisitos, análisis, diseño, construcción e

implantación, pruebas y mantenimiento. NDT-Suite está integrada por los diversos componentes, entre ellos, NDT-Profile, NDT-Quality y NDT-Driver.

NDT-Profile es una herramienta diseñada sobre un perfil definido, en base a los metamodelos de NDT, sobre la herramienta Enterprise Architect. El profile (perfil) sobre Enterprise ofrece una serie de herramientas y definición de artefactos propios para trabajar con la metodología NDT permitiendo una sencilla gestión de documentación.

Las fases de desarrollo incluidas en el proyecto se identifican por las siguientes siglas:

- EVS: documento del estudio de viabilidad del sistema.
- DRS: documento de requisitos del sistema.
- DAS: documento de análisis de sistema.
- DDS: documento de diseño del sistema.
- DPS: documento de plan de pruebas del sistema.
- DMS: documento de mantenimiento del sistema.

Además, se introduce una serie información adicional sobre el proyecto: Participantes: se describen las empresas y personas que participarán en el proyecto, Control de Versiones: se describen las diferentes líneas bases, y Objetivos del Proyecto: se describen los objetivos a cumplir en el proyecto.

3 Metodología

3.1 Descripción

La asignatura Taller de Programación I, de la carrera Licenciatura en Sistemas de Información, tiene como objetivo profundizar el estudio de herramientas de desarrollo de software orientadas a la plataforma web mediante la programación de aplicaciones. Busca ofrecer al alumno una visión amplia de las tecnologías utilizadas en el desarrollo de aplicaciones web, partiendo desde el diseño de páginas estáticas y de las tecnologías orientadas a la presentación (CSS, JavaScript), repasando tecnologías de cliente para mostrar luego tecnologías de programación para servidores, completando el recorrido con una visión general de acceso a base de datos.

En este proceso se pretende consolidar en el alumno las competencias requeridas para un analista programador, tales como el modelado y los métodos y herramientas para la especificación, diseño, implementación y evaluación de aplicaciones informáticas.

Esta asignatura contribuye específicamente a la formación del Analista Programador Universitario, título intermedio de la carrera, cuyo perfil comprende el desarrollo, modificación y mantenimiento de aplicaciones informáticas, mediante la utilización de herramientas de desarrollo de uso generalizado en el mercado laboral.

Por tanto, se espera que el alumno adquiera destrezas en programación mediante una intensa tarea de desarrollo siguiendo todas las etapas conceptuales de un proyecto de software, desde su especificación hasta su verificación y validación, incorporando además las buenas prácticas que exige la Ingeniería del Software para lograr desarrollos y soluciones cada vez más completas y robustas, haciendo énfasis en la documentación.

Para cumplir con este objetivo, se planteó la utilización de la metodología NDT enmarcada en el paradigma de Ingeniería Web guiada por modelos [7].

El eje del trabajo de la asignatura lo constituye el desarrollo individual de una aplicación web sencilla pero completa, que incluya todos los componentes necesarios: modelado de la aplicación, diseño gráfico y de contenidos, gestor de base de datos, tecnologías de programación en cliente y en servidor.

Cabe aclarar que, paralelamente al dictado de esta asignatura, los alumnos cursan la asignatura Ingeniería de Software I, con la cual se articulan conceptos, entre ellos, las técnicas de elicitación de requerimientos.

Para lograr los objetivos propuestos se realizaron las siguientes actividades:

1. Repaso de las técnicas de elicitación de requerimientos. Estas técnicas que se muestran en la Tabla 1, forman parte de los contenidos de la asignatura Ingeniería de Software I por tanto el repaso consistió en una breve reseña de las principales características de cada una de ellas para que los alumnos las tuvieran en cuenta para al especificar los requerimientos de su aplicación.

Tabla 1. Detalle de técnicas de elicitación de requerimientos

Fases	Actividades	Técnicas
Ingeniería de Requisitos	Obtener información sobre el entorno y definir objetivos	Entrevistas
		JAD
		Brainstorming
		Revisiones y búsqueda de información
		Cuestionarios
		Concept mapping
		Patrón para la definición de objetivos
	Identificar y definir los requisitos de almacenamiento de información	Patrón para la definición de requisitos de almacenamiento de información.
		Patrón para la definición de las nuevas naturalezas
	Identificar y definir los actores	Patrón para la definición de actores básicos
		Diagramas de representación de actores generalizados
		Matriz para la definición de incompatibilidad de actores
		Matriz para la definición de actores generalizados
	Identificar y definir los requisitos funcionales	Diagramas de casos de uso
		Patrón para la definición de los requisitos funcionales
	Identificar y definir los requisitos de interacción	Patrón para la definición de frases
		Patrón para la definición de prototipos de visualización
	Identificar y definir los requisitos no funcionales	Patrón para la definición de requisitos no funcionales

2. Capacitación en la metodología NDT. Se presentó a los alumnos la metodología de desarrollo NDT y se les instruyó en el uso de la herramienta NDT SUITE, que implementa dicha metodología.
3. Consignas para el desarrollo de una aplicación web. Se propuso a los alumnos el desarrollo de una aplicación web con la mayor cantidad de funcionalidades posibles. En primer lugar debían realizar la especificación de requerimientos de su

aplicación web utilizando los patrones aportados por la metodología NDT o el estándar IEEE 830-1998.

4. Definición de la primera instancia evaluativa: Esta consistió en la elaboración y entrega de la especificación de requerimientos de la aplicación a desarrollar.
5. Definición de la segunda instancia evaluativa: Consistió en completar el desarrollo de la aplicación, incorporando el acceso a una base de datos, entregando el producto final junto con su documentación. La utilización de la herramienta NDT Suite, se propuso en forma optativa.
6. Análisis de los trabajos presentados por los alumnos. Se analizaron 17 trabajos de los alumnos que utilizaron la metodología NDT, con el objetivo de evaluar el grado de aplicación de la metodología.
7. Encuesta de satisfacción a los alumnos. Se realizó una encuesta on line con la herramienta GoogleDocs, destinado a recabar la percepción de los alumnos respecto a la metodología de desarrollo propuesta.
8. Elaboración de Conclusiones

3.2 Análisis de los trabajos

De la primera instancia evaluativa se analizó el uso de las técnicas de elicitación utilizadas, cuya frecuencia de uso se muestra en la Tabla 2.

Tabla 2. Frecuencia de uso de técnicas de elicitación de requerimientos

Fases	Actividades	Técnicas	Nº de casos	%
Ingeniería de Requisitos	Obtener información sobre el entorno y definir objetivos	Revisiones y búsqueda de información anterior	2	11,76%
		Patrón para la definición de objetivos	15	88,23%
	Identificar y definir los requisitos de almacenamiento de información	Patrón para la definición de requisitos de almacenamiento de información	12	70,58%
		Patrón para la definición de las nuevas naturalezas	4	23,5%
	Identificar y definir los actores	Patrón para la definición de actores básicos	4	23,5%
		Diagramas de representación de actores generalizados	13	76,47%
	Identificar y definir los requisitos funcionales	Diagramas de casos de uso	13	76,47%
		Patrón para la definición de los requisitos funcionales	7	41,17%
	Identificar y definir los requisitos de interacción	Patrón para la definición de frases	3	17,64%
	Identificar y definir los requisitos no funcionales	Patrón para la definición de requisitos no funcionales	2	11,76%

Se observó la utilización de:

- Patrones de definición de objetivos: la mayoría (88%) de los alumnos utilizó esta técnica, la cual resulta especialmente útil para la comunicación con el usuario.

- Patrón para la definición de requisitos de almacenamiento de información: el 70,58 % de los alumnos lo utilizó, permitiendo la correcta definición de la estructura de la base de datos, que facilita la implementación.
- Diagramas de casos de uso y diagramas de representación de actores: 76,47 %

Las demás características proporcionadas por la herramienta no fueron aprovechadas.

En esta primera instancia de evaluación se pudo observar que la mayoría de los alumnos utilizó las características básicas de la herramienta.

Desde la percepción de los docentes de la asignatura el cumplimiento de la consigna no implicó mayor dificultad para los alumnos.

En la segunda etapa evaluativa, de acuerdo a los productos entregados y sus respectivas documentaciones generadas por la herramienta NDT-SUITE, se analizaron diferentes aspectos de la metodología:

a) Tipos de requerimientos: la Tabla 3 muestra los tipos de requerimientos especificados en cada uno de los trabajos.

Tabla 3. Tipos de requerimientos especificados

Tipos de Requerimientos	**Trabajos de los alumnos**																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Almacenamiento de datos	D	D	I	I	D	D	D	D	D	ND	D	D	D	D	D	ND	D
De actores	D	D	I	D	ND	D	D	D	D	D	D	D	D	D	D	D	D
Funcionales	D	D	I	I	D	D	D	D	D	ND	D	D	D	D	D	D	D
De interacción	ND	I	ND	ND	ND	ND	ND	ND	ND	ND	D	ND	D	ND	ND	ND	ND
No funcionales	ND	ND	ND	ND	ND	ND	ND	ND	ND	D	ND	ND	ND	ND	D	ND	D

D Definidos
ND No Definidos
I Incompletos

Se observa que la mayoría pudo definir los requerimientos de almacenamiento, de actores y los funcionales.

Muy pocos definieron los requerimientos de interacción, como así también, muy pocos especificaron los requerimientos no funcionales.

b) Información adicional sobre el proyecto. En la Tabla 4 se muestran los distintos ítems de información adicional que fueron utilizados por los alumnos.

Tabla 4. Información adicional del proyecto.

Información adicional	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Participantes	D	D	D	D	D	D	D	D	D	ND	D	D	D	D	D	D	D
Control de Versiones	-	-	-	-	-	-	-	-	D	-	-	-	-	-	-	D	-
Objetivos del proyecto	D	ND	D	D	D	D	ND	D	ND	ND	ND	D	D	D	D	D	D

De este punto de análisis se desprende que la mayoría de los alumnos pudieron definir los participantes involucrados en el desarrollo del proyecto, pero la mayoría no

registró los cambios en el Control de Versiones, aun cuando presentaron más de una versión del producto.

En cuanto a los Objetivos del proyecto, se observa que algunos alumnos tuvieron dificultades para definir los objetivos del proyecto.

c) Análisis de la especificación de los requerimientos en NDT. En la Tabla 5 se observan los distintos ítems analizados en las especificaciones

De los valores mostrados en la tabla 5, se desprende que no hubo dificultad en los alumnos en definir con claridad los objetivos de la aplicación web.

Tabla 5. Análisis de especificación de requerimientos en NDT

Casos	Claridad en la definición de los objetivos	Coherencia en la definición de los datos de almacenamiento	Completitud en los casos de uso	Consistencia entre los términos utilizados en las diferentes descripciones	Consistencia con el problema (se documenta correctamente con el Caso de Uso)
1	B	MB	B	B	MB
2	MB	B	B	B	B
3	MB	MB	MB	MB	MB
4	MB	R	R	R	R
5	MB	MB	MB	MB	MB
6	B	MB	B	B	B
7	B	B	MB	MB	MB
8	MB	B	MB	MB	MB
9	MB	MB	MB	MB	MB
10	MB	B	R	B	B
11	B	MB	MB	MB	B
12	MB	MB	MB	MB	MB
13	B	MB	B	B	B
14	MB	MB	MB	MB	MB
15	B	B	MB	MB	B
16	MB	R	MB	MB	MB
17	MB	MB	MB	MB	MB

Completar los casos de uso no representó dificultades en las especificaciones, como tampoco se observó inconsistencias en documentarlos correctamente y en el uso de términos distintos a lo largo del trabajo. Los alumnos tienen muy presentes esos conceptos porque se encuentran cursando la asignatura Ingeniería de Software I.

En el análisis de los productos finales se pudo comprobar que, si bien algunos alumnos describieron requerimientos de datos que luego no utilizaron, o viceversa, la herramienta les sirvió para detectar esta situación.

3.3 Análisis de satisfacción de los alumnos

Para evaluar el grado de satisfacción de los alumnos respecto a la metodología de enseñanza, se diseñó un cuestionario on-line que se envió a los alumnos. Dado que el cursado ya había finalizado, solamente 7 alumnos aportaron sus opiniones, que se consideran ilustrativas de la situación.

Las cuestiones abordadas fueron las siguientes:

1. Simplicidad de la herramienta
2. Claridad de los distintos módulos de la herramienta

3. Utilidad del manual de usuario
4. ¿La herramienta facilitó el desarrollo de la aplicación?
5. Importancia de la documentación en el desarrollo de software
6. ¿La herramienta contribuyó a la documentación de la aplicación desarrollada?
7. ¿Volvería a utilizar la herramienta en el desarrollo de una nueva aplicación?

Los resultados se presentan gráficamente en las figuras 2 a 7.

Figura 2. Simplicidad de la herramienta

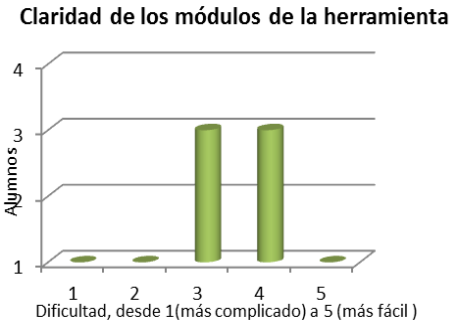


Figura 3. Claridad de los módulos

Figura 4. Utilidad del manual de usuario

Figura 5. Desarrollo de la aplicación

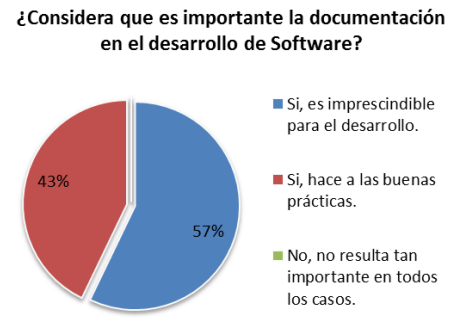


Figura 6. Importancia de la documentación

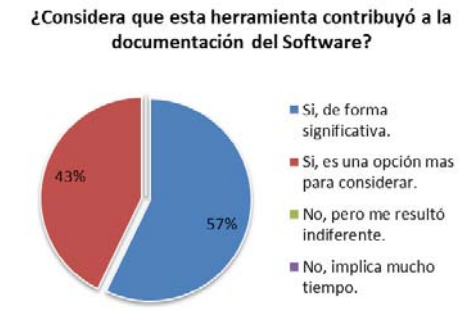


Figura 7. Contribución a la herramienta a la documentación

En cuanto a la consulta de si volverían a utilizar la herramienta en el desarrollo de una nueva aplicación, la totalidad de los alumnos contestó en forma positiva.

Analizados los resultados obtenidos, se observa que en cuanto a la facilidad de uso de la herramienta y de la comprensión de los distintos módulos (Fig. 2 y 3), la valoración fue mayoritariamente entre 3 y 4 para una escala de 1 a 5, donde 1 es más difícil y 5 más fácil, lo que significa que la comprensión fue de una dificultad media,
En cuanto a la utilidad del manual de la herramienta (Fig. 4), el 43% dice que lo utilizó permanentemente, un 28% lo leyó pero no lo necesitó y un 29% opinó que el manual no era muy claro. Vinculado con la herramienta, se les consultó también si su utilización facilitó el desarrollo de la aplicación (Fig. 5), la mayoría (72%) contestó que sí, pero que no fue imprescindible su uso.
Respecto de la importancia de la documentación en el desarrollo y de si la herramienta NDT contribuyó a la documentación de la aplicación desarrollada (Fig. 6 y 7), mayoritariamente (57% en ambos casos), opinaron favorablemente.
De las respuestas de los alumnos se puede inferir que apreciaron la herramienta por su aporte a la documentación pero, en esta etapa formativa, no les resultó imprescindible. Esta devolución servirá para revisar las consignas dadas en esta estrategia de enseñanza, buscando una mayor integración del modelado de la aplicación, utilizando la herramienta NDT-Suite, y el desarrollo de la misma.

4 Conclusiones

Este trabajo presenta los resultados de una estrategia de enseñanza que incorpora la herramienta NDT-Suite para el desarrollo de aplicaciones web.

La aplicación de NDT a proyectos reales resultó enriquecedora. En este caso, ha sido especialmente interesante porque los alumnos comprobaron que siguiendo las buenas prácticas en el desarrollo de software proporcionadas por la herramienta NDT Suite se obtiene un producto de calidad.

Tanto de los trabajos realizados por los alumnos como de las opiniones de los mismos se desprende que son conscientes de la importancia de seguir una metodología de desarrollo de software para obtener un producto de calidad. La herramienta contribuyó a la valoración de la documentación en los proyectos de software.

Al ser una herramienta muy completa, los alumnos podrán seguir avanzando en la utilización de los demás componentes, como por ejemplo el NDT Quality, que comprueba criterios de calidad.

Como trabajos futuros, además de resaltar las futuras aplicaciones pero ya en otros perfiles del NDT-Suite, se pretende ampliar la estrategia incorporando la utilización de la herramienta en otras asignaturas vinculadas con la Ingeniería del Software.

Referencias

1. Abrahão S.M., Poels G., Pastor O., “Evaluating a Functional Size Measurement Method for Web Applications: An Empirical Analysis”, in Proceedings of Tenth International Software Metrics Symposium (METRICS’04), Chicago, Illinois, USA, pp. 358-369, 2004.
2. Pedrozo Petrazzini, Osmar G., Medina, Yanina, Dapozo, Gladys Noemí. “Análisis Comparativo de Metodologías Web”. XIX Reunión de Comunicaciones Científicas y Tecnológicas. Edición 2013. Universidad Nacional del Nordeste. Resistencia. Chaco. Disponible en: http://www.unne.edu.ar/trabajando/com2013/exactas.php

3. Escalona, M. J. and Koch, N. 2007. Metamodelling the requirements of Web systems. Lecture Notes in Bussiness Information Process, vol. 1, Springer Verlag, 267–288.
4. Escalona, M.J., Mejías, M., Torres, J. "Methodologies to develop web information systems and comparative analysis". UPGRADE.TVol. III, No. 3, Junio 2002.
5. NDT (Navigational Development Techniques). Metodología desarrollada por el Grupo de Investigación Ingeniería Web y Testing Temprano (IWT2). Universidad de Sevilla. Disponible en: http://www.iwt2.org/web/opencms/IWT2/inicio/?locale=es
6. Escalona, M.J., Torres, J., Mejías, M. (2002). "Requirements capture workflow in Global Information Systems". Proceedings of OOIS. Springer-Verlag. Montpellier, France.
7. Escalona, M.J.,Gutierrez J.J., Ortega J.A., Aragón G., Pérez Pérez M., Ponce J., "NDT-Suite, una solución práctica para el uso de NDT".
8. Escalona, M.J., Aragón. G."NDT. A model-driven approach for web requirements". IEEE Transaction on Software Engineering. Vol. 34. Nº3. Mayo/junio 2008. IEEE Computer Society.

III ETHICOMP LATINOAMÉRICA

- ETH -

CACIC2013

III ETHICOMP LATINOAMÉRICA

- ETH -

ID	Trabajo	Autores
5711	**Una conversación con Matthew Sher sobre privacidad y la amistad**	Maria Beatriz García (UNLP), William M. Fleischman, (Villanova University, USA), Joaquin Bogado (UNLP)
5705	**Complexity is Free, but at What Cost? A Survey of the Current Uses of 3D Printers and the Ethical Concerns that Will Arise from Their Continued Use**	Kelly Gremban (Villanova University, USA)
5853	**Greedy for Green: The Motivations and Hindrances of Maintaining Environmental Sustainability in the Commercial Industry of Cloud Computing**	Bradley S. Cantor (Villanova University, USA)
5685	**Why We Should Not Build Autonomous Robotic Weapons**	William M. Fleischman (Villanova University, USA)

Una conversación con Matthew Sher sobre privacidad y la amistad

María Beatriz García
Psicoanalista
mariabeatriz.garcia @ gmail.com

William M. Fleischman
Departments of Computing Sciences and Mathematical Sciences
Villanova University
Villanova, Pennsylvania 19085, U. S. A.
william.fleischman@villanova.edu

Joaquín Bogado
LINTI – UNLP
jbogado@linti.unlp.edu.ar

Abstract: El siguiente trabajo indaga en la importancia de la privacidad para la dignidad humana desde dos miradas esenciales al tema: ética y psicológica, teniendo cuenta la privacidad en las redes sociales y el comportamiento de la generación del milenio según el artículo de Matthew Sher al respecto. Partiendo de la necesidad y la valoración de la de privacidad se va desarrollando su conceptualización a la luz de la Etica Nicomaquea, su lugar en la estructuración del psiquismo y su impronta en la actualidad, donde parece estar amenazada por las nuevas tecnologías y el uso de las redes sociales por los jóvenes de la hipermodernidad.

Keywords: Privacidad, Amistad, Redes sociales, la Ética Nicomaquéa, la Ética en relación con la informática

1 Introducción

En un perspicaz artículo presentado al workshop II ETHICOMP Latinoamérica en 2012, Matthew Sher se examinó la facilidad con la que la generación del milenio ha adoptado los medios sociales y la tecnología de la internet. [1] Habló de la comodidad que sienten los jóvenes en la negociación de las políticas de privacidad de Facebook, Google y otros medios de comunicación social, la conciencia de las

vulnerabilidades asociadas con la pérdida de la privacidad y su decisión en gran parte consciente de compartir información personal a cambio de los beneficios percibidos de la autoexpresión y la actividad comercial. Sher se refirió a las afirmaciones del CEO de Facebook Mark Zuckerberg en el sentido de que "la edad de privacidad ha terminado" y que "las personas realmente han logrado sentirse cómodas no sólo con compartir más información de diferentes clases, sino también mas abiertamente y con más gente. Esa norma social es algo que ha evolucionado con el tiempo."

Una de las consecuencias inquietantes de esta tendencia es que reduce el concepto de privacidad a una mercancía que puede canjearse por acceso comercial y estima personal. De hecho, en un artículo recién publicado sobre la privacidad y la ley, Benjamin Wittes explora "la posibilidad que el avance de la tecnología y la proliferación de datos personales en manos de terceros nos ha dejado con un debate conceptualmente anticuado, cuya dependencia del concepto la privacidad no resulta en una guía útil en las cuestiones de política pública que enfrentamos". [2].

En este trabajo, queremos rehabilitar la importancia del valor de la privacidad y, dibujando las conexiones al concepto de la amistad, señalar ciertos daños asociados con el comercio libre y no reflexivo de la información privada y personal.

2 ¿Por qué necesitamos y valoramos la intimidad?

Según el análisis cuidadoso de Ruth Gavison [3], privacidad tiene un papel importante en la promoción de valores como la libertad, autonomía, autenticidad y salud mental que consideramos esenciales para llevar una vida productiva y satisfactoria. Podemos enumerar algunas de estas funciones:

- Es un espacio libre de distracción para que nos podamos concentrarnos en nuestros propios pensamientos
- Es un espacio en el cual podemos explorar, de manera provisional, ideas y posibilidades sensibles
- Es un espacio para el desarrollo de las relaciones íntimas
- Es un espacio libre de la presión y la influencia de los demás, por ejemplo
    - Gobiernos que deseen dirigir o inhibir nuestra capacidad de organizarse políticamente
    - Aquellos en una posición de autoridad en cuanto a nuestro trabajo y actividades profesionales
    - Los padres y otros miembros de la familia para que seamos capaces de desarrollar nuestra propia identidad auténtica

Edward Bloustein afirma que en el caso extremo de la pérdida de privacidad, "El hombre que se ve obligado a vivir cada minuto de su vida, entre otros, y que cada una de las necesidades, el pensamiento, el deseo, la fantasía o la gratificación está sometida al escrutinio público, ha sido privado de su individualidad y dignidad humana. Tal individuo se fusiona con la masa. Sus opiniones, siendo públicas, nunca tienden a ser diferentes; sus aspiraciones, siendo conocidas, siempre tienden a ser convencionalmente aceptables; sus sentimientos, siendo exhibidos abiertamente, tienden a perder su calidad de ardor personal único y convertirse en los sentimientos de cada uno. Tal ser, aunque inteligente, es fungible; no es un individuo." [4].

¿Cuáles son los impedimentos para el ejercicio saludable de privacidad? En primer lugar, podemos destacar el constante canto de sirena del correo electrónico y los medios de comunicación social. Además podemos subrayar el impulso a usar estos medios para exponer incluso los más íntimos pensamientos y actos.

La felicidad es la respuesta a la cuestión de la finalidad de la vida humana. Los avances científicos y tecnológicos actuales no siempre se nos presentan como progreso pero sin duda apuntan conscientemente al "bien-estar", avatar de la felicidad reducida a los bienes que cada uno posee. En la sociedad actual tenemos, por un lado, la profusión de los objetos de las tecno-ciencias incidiendo sobre los modos de goce de la familia, aunque más no sea por la influencia que tiene el reparto para el uso de los "gadgets" que se ponen en juego en su vida cotidiana. Por otro, encontramos el ámbito de lo público donde, entre otros fenómenos, se observa un empuje a decirlo todo, a contar todo en todas partes, a ver y ser visto, y especialmente en los medios de comunicación que vuelven todo mercancía. A traves de ellos, hoy la vida privada con sus modos de goce, intenta hacerse pública.Se trata de la promoción de un espectáculo donde la relación social entre personas queda mediatizada por la imagen, pero también del ejercicio de una violencia que implica el asesinato de la singularidad del sujeto, puesto que, capturado por un espectáculo al estilo de los reality shows, en apariencia participa de algo que parece muy subjetivo, pero en verdad queda reducido a una simple imagen.La mirada promovida es una mirada sin vergüenza, una mirada que no se erige en una instancia que juzga, sino que se reduce a otro que también goza. Es la época de la transparencia donde no existe lo invisible. La "era de la privacidad" ha terminado, según declara el fundador de Facebook. Por otro lado también parecen desfallecer los secretos: Julian Assange dijo que también había terminado el tiempo de los secretos de Estado.

3 Sobre la vida privada

Comenzamos por valernos del equívoco que introduce este título, poniendo en consideración sus dos acepciones posibles. Nuestro término "privado" procede del latín. Y en esta lengua "privatus" es el simple ciudadano particular en su calidad de tal: mientras que la "privatio" tiene el sentido, que resuena a veces en algunos usos de nuestra palabra, de la carencia.

Si queremos plantear soluciones a los problemas a los que nos enfretamos hoy día, deberíamos volver a los puntos centrales de las teorías éticas clásicas, que han quedado olvidados en las teorias políticas y morales actuales. La concepción como esferas separadas de lo público y lo privado no siempre ha existido, sino que ha evolucionado a lo largo de la historia, en la mente de los teóricos, así como en la conciencia de los ciudadanos. Separar tajantemente lo público y lo privado como en la actualidad parece inevitable, supone un error de base.

Entonces cabría preguntarse: ¿Hay algo en la filosofía aristotélica equiparable a nuestra comprensión de la privacidad?

4 Un retorno a Aristóteles

En la Etica Nicomaquea y en la Política de Aristóteles [5], las relaciones humanas y la eticidad del Estado son cuestiones relevantes. En nuestra sociedad individualista y fragmentaria parece haberse olvidado lo que ya Aristóteles afirmaba: que el hombre es un animal que sólo puede vivir relacionándose con los demás.

En la Grecia Clásica se entendía por "público" a todo lo relacionado con la polis: tanto el individuo en su condición de ciudadano, de extranjero, de exiliado; como todo lo que tiene que ver con el gobierno común y la preocupación por los asuntos políticos;el sistema de las instituciones,así como las leyes y las normas relativas al ciudadano. Es este ámbito así entendido un reino de la igualdad.

Por privado, en cambio, entendían todo lo concerniente al individuo en el terreno doméstico: el ser humano en sus relaciones con otros individuos; en sus roles de esposo, esposa, padre, madre, hijo; pero también todo lo que tiene que ver con las posesiones privadas, los intereses, necesidades, aspiraciones, deseos y derechos de los individuos en el oikos. Ambito este, entonces, de la desigualdad y la diversidad. También era éste el ámbito de lo íntimo.

La comprensión de la relación que hay entre la filosofía práctica y teórica en el conjunto del cuerpo filosofal aristotélico es lo que nos permitirá entender el sentido de la relación entre lo público y lo privado, entre la ética y la política.

Siendo la naturaleza humana la materia prima sobre la que deben fundarse las reglas éticas y políticas, será necesario visualizar la teoría del hombre que tiene el Estagirita; concepción que marcará las teorías ético-politicas.

Como es bien sabido, es precisamente el logos el rasgo definitorio del hombre aristotélico, lo que hace aparecer la posibilidad de la comunicación entre los hombres y la posibilidad de no conformarnos con el simple vivir y ,por tanto, la necesidad de la política.

La razón de ser del mundo y del hombre aristotélico atiende a un telos que está definido de modo natural, fin al que debemos acercarnos para perfeccionarnos y cumplir así nuestra función específica de hombres. Esta función, que Aristóteles llamará Bien, es una función o trabajo que se presenta en tanto el hombre no es sólo logos, sino un compuesto de logos y deseos. Pero, lo que realmente condiciona la ética y la política es que lo determinante en el hombre es su logos que hace de él un ser relacional que no puede vivir en soledad. Esto lleva inevitablemente a atender la convivencia, la comunicación y hacer de la ética y la política un cuerpo unido, soldado.

La búsqueda del Bien conllevan dos caminos distintos pero complementarios: por un lado el hombre busca su eudaimonia desde su automovimiento, siendo él mismo quien le da intencionalidad a dicho movimiento. Este camino será estudiado por la ética. Por otro lado, el hombre puede y debe buscar su felicidad en conjunción con los otros hombres, aunando principios comunes voluntarios. Movimiento común que será estudiado por la política.

Desde el comienzo de la historia, los hombres han luchado, empírica o intelectualmente, frente a la coacción para mantener o ganar su libertad. En la política actual éste sigue siendo un problema recurrente. Lo público y lo privado tienen así una relación estrecha con la libertad y dependiendo de cual sea nuestra

concepción del par público-privado y de la valoración que de uno u otro hagamos, así daremos lugar de una u otra manera a la libertad y así la defenderemos. Y viceversa: nuestra concepción de la libertad dará valor a uno u a otro aspecto de nuestra vida, al público o al privado.

Podemos entender estos conceptos de modos diversos, así como referirnos a ellos con distintos nombres, pero no hay duda de que son nociones que guían y han guiado las reflexiones políticas y éticas de todos los tiempos. Pueden aparecer con el nombre de poder frente a la conciencia, de lo oficial frente a la resistencia individual o como la ley frente a la voluntad individual, como el Estado frente a la persona, la casa frente a la ciudad, el poder público frente a los derechos individuales, la solidaridad frente al individualismo o como se nos presenta en el mundo moderno: el hombre público frente al sujeto privado.

Tampoco hay dudas que los conceptos de público y privado son categorías que aparecieron en la cultura griega clásica, polaridad que está presente ya en la literatura griega antigua, explícitamente en la Odisea, pero que fue más elaborada en el período democrático de la Atenas clásica.

Es importante recalcar que esta polaridad no significaba en el mundo griego una separación de esferas. En la vida de la polis la realidad es una estrecha relación entre lo público y lo privado, y si bien es cierto que podían entrar en conflicto, es, precisamente, la reciprocidad de ambas lo que permitía que la vida privada del hombre y la vida pública o de la polis, fueran ambas civilizadas y humanas; pues, los actos y las decisiones de uno y otro ámbito no estaban aisladas de sus consecuencias en el ámbito contrario.

Es la polis con sus leyes la condición de posibilidad de la vida privada del individuo, siendo el hombre comúnmente aceptado en la cultura griega como un ser que tiende a la comunidad, no sólo por interés propio, sino también por naturaleza o necesidad. Y aún más, porque ambos, polis e individuo, comparten unos fines comunes.

Hoy dia hemos convertido en necesaria y evidente una separación que no tenía cabida en el mundo aristotélico.

En las sociedades actuales, modernas,hemos pasado a una forma de democracia representativa en la que esperamos participar lo menos posible en los asuntos públicos para tener más tiempo libre para nuestros asuntos privados. Nos hemos olvidado de la concepción del hombre aristotélico donde sólo es tal en tanto se vive en comunidad, en una comunidad bien organizada, una comunidad bajo la ley, expresión de la justa razón. Para un griego del siglo IV, obrar según la ley era obrar según la razón. El ciudadano griego se identifica con esas leyes porque sabe perfectamente que sin leyes no hay libertad posible: volvería a la tiranía o a la oligarquía. Las leyes son entonces el garante de su autonomía, no un obstáculo a su vida privada. La libertad individual se afirmaba en el tomar parte en el poder colectivo.

Tras la aceptación general de que la libertad pública y la libertad privada no son sólo diferentes sino opuestas y excluyentes una de otra, ha queda eclipsada la concepción misma de la democracia ateniense. Es notable que nosotros ya no podamos disfrutar de la libertad de los antiguos, las razones son sobre todo nuestra concepción de lo público y lo privado.

Para nosotros lo público ha dejado de ser garante de nuestra vida privada; la política está divorciada de la ética y sin duda, las leyes han dejado de ser expresión de la pura razón. No es el bien común, sino el interés privado, lo que prima y ambos son entendidos como incompatibles. La tan mentada globalización nos ha hecho perder la conciencia de pertenecer a una comunidad, como vivenciaba el hombre de la Grecia Clásica.

Esas son las verdaderas razones de que no seamos capaces de disfrutar de la libertad antigua.

5 Esencial-mente privados

Se denomina razón a la introducción de un orden de determinaciones en la existencia humana, en el orden del sentido. El descubrimiento de Freud es el re-descubrimiento, en un terreno virgen, de la razón.

Ese terreno virgen que, ya está presente en la Etica aristotélica y que corresponde a la parte "irracional" del alma humana, Freud la denomina y sistematiza como Lo Inconsciente. Es lo más íntimo, lo más privado y a la vez lo más ajeno a nosotros mismos. Es esa intimidad que nos es éxtima, desconocida al yo de la conciencia.

El término "mente" se ha equiparado al término "psique" y desde Aristóteles a nuestros días son inumerables las corrientes que han tratado de explicar el funcionamiento psíquico.

Pasar del campo de la Filosofía al campo de la Psicología implica dar un salto cualitativo. Si como pudimos ver es la ética y la política en el mundo clásico lo que nos permite, localizando al ciudadano, articular en un sentido los conceptos de lo privado y lo público; será, lo privado como acepción en el sentido de "carencia", de "falta", lo que nos permitirá articular lo que del psiquismo está en juego en el fenómeno que nos interpela: la aparente despreocupación de los sujetos modernos por el cuidado de su privacidad. Despreocupación que parecería reflejarse en su máxima expresión en las denominadas redes sociales. Pasamos así de la consideración de lo público y lo privado en una esfera que podríamos decir de una realidad general a una realidad particular, subjetiva…

No hay "realidad", para el ser humano, en el sentido de una experiencia inmediata o no mediada. Esta mediación el hombre la lleva a cabo a través de su psiquis. Hay una realidad externa y una realidad interna. Mundo interno,yoico,que le permitirá captar, interpretar y relacionarse con la realidad externa. La noción de realidad está articulada mediante la significación, el mundo simbólico del sujeto y la esquematización característica de las imágenes que constituyen su mundo imaginario.

Mundo interno, acaso, máxima expresión de lo privado. Privatio, privado, carente… falta, constitutiva y constituyente del sujeto. Falta que en el psiquismo se da en tres registros: imaginario, simbólico y real, estructura triádica borromea del ser.

El sujeto del Inconsciente no es un sujeto dado sino que es un sujeto que se estructura. La función que para el hombre desempeña la imagen de su propio cuerpo es fundamental en la estructuración psíquica. La imagen especular, imagen virtual, es constitutiva del Yo, es su reflejo, su imagen anticipada, lo que constituirá su ser (registro imaginario). Relación muy estrecha con la superficie del cuerpo en tanto

reflejada en una forma. No se trata de la superficie sensible, sensorial, sino de esa superficie en tanto está reflejada en una forma. La imagen de la forma del otro es asumida por el sujeto.

La imagen estará siempre presente en las diferentes etapas del sujeto. La veremos aparecer ya sea como velo, como pantalla que muestra, que vela la falta, ya sea como espectáculo que llama a ver, o como producción artística, provocando la mirada….la imagen como respuesta al vacío, a la falta. Ser de semblante, "identificación imaginaria": lo que estaba afuera se convierte en el adentro.

El hombre se aprehende como cuerpo, como forma vacía del cuerpo, en un movimiento de báscula, de intercambio con el otro. Aprenderá a reconocer invertido en el otro todo lo que en él está entonces en estado de puro deseo, deseo originario, inconstituído y confuso, deseo que se expresa en el gemido o llanto del recién nacido. Aprenderá cuando se ponga en juego la comunicación. Esta anterioridad no es cronológica sino lógica. Antes que el deseo aprenda a reconocerse por el símbolo, sólo es visto en el otro. En el origen, antes del lenguaje, el deseo sólo existe proyectado, alienado en el otro. La tensión que produce no tiene salida….más que la destrucción del otro.

Ahora bien, ¿qué significa decir "yo"?. ¿Significa acaso lo mismo que capturarse siendo en una imagen especular?

Hay en francés pero no en castellano dos modos del pronombre de primera personal del singular, que dan cuenta de la incidencia del lenguaje en la estructuración psíquica: "moi" y "je."

Yo (Je) es un término verbal cuyo empleo es aprehendido en una cierta referencia al otro, referencia que es una referencia hablada. El Yo (Je) nace en referencia al Tú, en una relación donde el otro le manifiesta órdenes, deseos, que él debe reconocer. El sujeto está entonces en el mundo del símbolo, es decir en un mundo de otros que hablan. Su deseo puede ser dicho, puede pasar entonces por la mediación del reconocimiento. El grito se convierte en llamada, estructura de discurso que precipita el lazo social. La inscripción del sujeto en el orden simbólico es la posibilidad misma de la pacificación de la agresividad emanada de lo imaginario.

Mundo simbólico que garantiza un ordenamiento en la relación al otro.

Esta es la vía por donde el niño aprende el orden simbólico y accede a su fundamento: la ley. Ley que impone un límite a través de una prohibición que inscribe una falta estructural en el psiquismo. "Falta en ser" que instituye un lugar para el sujeto.

Dijimos que en el ser humano la realidad no es aprehendida de manera inmediata y debemos agregar que tampoco es posible un dominio pleno de la realidad, algo se escapa, volviendo siempre al mismo lugar. Hay algo de "imposible" en la realidad para el ser humano. Ese imposible es lo Real, lo que no puede ser simbolizado. Siempre hay algo de impensable…avance de la ciencia tapando el agujero de lo Real, falla imposible de tapar. Impotencia del conocimiento que hoy muestra más que nunca, a pesar de los logros de la ciencia y la tecnología, ese Real, la felicidad, imposible de atrapar.

6 De la Philía Aristotélica a los amigos del Face…

El preámbulo de la Declaración Universal de los Derechos Humanos, proclamada por las Naciones Unidas, considera que "la libertad, la justicia y la paz en el mundo tienen por base el reconocimiento de la dignidad intrínseca y de los derechos iguales e inalienables de todos los miembros de la familia humana"; y en su artículo 12 deja expresamente plasmada la privacidad como uno de los derechos inalienables relacionados con la dignidad humana: "Nadie será objeto de injerencias arbitrarias en su vida privada, su familia, su domicilio o su correspondencia, ni de ataques a su honra o a su reputación. Toda persona tiene derecho a la protección de la ley contra tales injerencias o ataques."

En el artículo de Sher [1], se puntualiza que "The decision to share information online is a conscious choice, and most Millennials understand the tradeoff that comes with connectivity. In an April 2010 Harris poll , 85% of the surveyed Millennials acknowledged that, by participating in social media, they are giving up part of their privacy." Más adelante agrega, "…In spite of their awareness, Millennials apparently hold that the benefits of social

media greatly outweigh the privacy risks that come with it."

En el estudio "Los adolescentes y las redes sociales"[7], se consigna que lo que más valoran los jóvenes de sí mismos es su grado de popularidad... y qué necesita un adolescente para ser popular?... "amigos, humor y espontaneidad" es lo que respondieron a esta encuesta 3500 alumnos de escuelas secundarias de Argentina. Este mismo estudio en referencia a los riesgos en las redes sociales reflejó datos interesantes: 95% no cree en los riesgos de internet, 90% se siente inmune frente a lo que puedan encontrar, 75% cree en todo lo que dice la Red, 60% cree que sólo amigos ven su página personal y el 90% dice que en su casa no hay reglas de uso de la Internet. En sus propias palabras, dicen: "los riesgos son manejables", "me tengo confianza, soy hábil con la tecnología", "es más importante conocer gente, que pensar en los riesgos", "me gusta abrir mi página para que la vean todos", "no me imagino qué riesgos pueda tener estar en una red social", "la red social es la responsable de los riesgos y los tiene controlados. Tanto lo que consigna Sher en su artículo como lo reflejado por este estudio, marcan que los jóvenes sostienen que los beneficios de estar conectados a las redes sociales son mayores que sus posibles riesgos.

Virtualidad no es un término que pertenezca al campo conceptual del psicoanálisis ni tampoco al de la ética clásica, sin embargo, tanto un enfoque como el otro, nos pueden ayudar a comprender el por qué aquello de lo que se trata, está cada vez más omnipresente en el mundo, cambiando, perforando, subvertiendo?...conceptos como el de la privacidad, tan insoslayable cuando de la dignidad humana se trata.

Si en *Etica Nicomaquéa*, Aristóteles se ve llevado a dedicarle dos libros enteros al tema de la amistad, hemos de pensar que tamaño tratamiento, por sobre reflexiones éticas como la felicidad, la virtud, el placer, la justicia, no es algo azaroso.

Responde a la convicción aristotélica de que la amistad es algo especialmente valioso, diríamos que algo único, en la vida de los seres humanos. La amistad, en efecto, no es un aliciente más, entre otros, para una vida feliz: es --en palabras del propio Aristóteles-- "lo más necesario para la vida", lo más necesario para una vida

feliz. Por eso, dice Aristóteles, “nadie querría vivir sin amigos, aun estando en posesión de todos los otros bienes.”

En definitiva, puesto que el ser humano es un animal social, que naturalmente tiende a la convivencia con otros seres humanos, la amistad constituye la realización más plena de la sociabilidad y la forma más satisfactoria de convivencia.

Amistad se dice en griego philía, palabra de la misma raíz que el verbo phileîn, que significa “querer”. En griego, philía abarca todo tipo de relación o de comunidad basado en lazos de afecto, de cariño o amor, y de ahí que Aristóteles incluya, bajo esta denominación, relaciones tan dispares como el cariño entre padres e hijos, la relación apasionada entre amantes, la concordia civil entre conciudadanos, y la relación que nosotros consideramos más estrictamente como amistad.

Dos perspectivas sigue Aristóteles en el tratamiento del tema de la amistad : los distintos tipo de amistad y la amistad perfecta. En la primera línea de pensamiento, Aristóteles convencido de que la diversidad de opiniones es siempre una muestra de la complejidad del tema a tratar, desarrolla el hecho incuestionable de que existen opiniones muy diversas y contrapuestas acerca de la amistad.

No obstante las distintas opiniones Aristóteles encuentra que hay un núcleo común significativo en las distintas nociones de amistad. En primer lugar, la amistad se define por el querer, por el afecto. Ahora bien, no toda forma de querer es propiamente amistad. En efecto, la amistad exige un querer mutuo, recíproco y, además, que sea conocido y reconocido por ambos, por ambas partes. Si el querer no es recíproco, o si una o las dos partes desconocen la reciprocidad de su querer, no cabe hablar de amistad en sentido estricto.

Teniendo en cuenta las diversidades del querer , Aristóteles reconoce tres formas o tipos de amistad: la amistad basada en la utilidad, la amistad basada en el placer y la amistad basada en el bien, es decir, en la virtud o excelencia de la persona a la cual se quiere. En las dos primeras formas de amistad no se quiere al amigo por sí mismo, sino accidentalmente, no se quiere al amigo por lo que es o por el que es, sino porque coincide que tal individuo nos resulta útil o placentero.

Ahora bien, Aristóteles reflexiona sobre la amistad desde una perspectiva ética, desde la perspectiva concerniente a la felicidad, a la vida buena, digna y satisfactoria. Desde esta perspectiva, Aristóteles considera que las amistades basadas en la utilidad y en el placer son formas deficientes de amistad comparadas con la amistad basada en el bien, en la virtud, a la cual denomina amistad perfecta. En efecto, solamente en esta forma de amistad se da la benevolencia en sentido estricto, es decir, el querer al amigo y el querer el bien del amigo por él mismo, que es lo que define la auténtica amistad.

La amistad perfecta — por tanto, la amistad auténtica, la que merece tal nombre — es aquella que se basa en la excelencia, en la virtud, y en la cual el amigo es querido por sí mismo. Ahora bien, cabe preguntarse cuando Aristóteles dice que el amigo es querido por sí mismo, ¿qué entiende por “sí mismo”? ¿qué ha de entenderse que es el sí mismo del ser humano?

En un sentido el “sí mismo” de cada cual se manifiesta en el modo en que uno vive, en el modo en que uno realiza su propia existencia, en definitiva, en las acciones que uno lleva a cabo. Pero no en cualquier tipo de acciones, sino en las acciones o actos elegidos. Aristóteles distingue, en su ética, entre actos voluntarios y actos elegidos Actos voluntarios son aquellos que se realizan con conocimiento de lo

que se está haciendo y sin coacción alguna externa que fuerce al individuo a su realización.

No toda acción voluntaria es, sin embargo, una acción elegida. La elección comporta conocimiento racional, comporta deliberación, y Aristóteles la caracteriza como “inteligencia deseosa, o bien, deseo inteligente”. La elección es el principio propiamente humano de la acción, es aquel principio de donde surgen las acciones verdaderamente humanas.

Lo cual significa que, en último término, cada cual es responsable de su propio carácter ya que éste resulta, en último término, de nuestras propias elecciones.

Como toda disposición ética, la amistad se refiere primariamente a la elección, en este caso a la elección adecuada de los amigos. Y la elección adecuada del amigo es la elección del amigo que es bueno, que es excelente. De este modo puede decir Aristóteles que “al querer al amigo quieren su propio bien, puesto que cuando alguien bueno se convierte en amigo querido, se convierte en un bien para aquél que lo quiere. De modo que uno y otro quieren su propio bien, y se recompensan recíprocamente por igual.” La tesis de Aristóteles es que el amor al amigo constituye una extensión del amor a sí mismo. Entonces el amigo “tiene para con el amigo la misma disposición que para consigo mismo.”

Llegado a este punto de la reflexión se impone una pregunta: ¿qué causará que, después de más de 2500 años, siga siendo la amistad algo tan imprescindible para el ser humano? Una respuesta posible llega del lado de lo ya dicho sobre la estructuración psíquica. Existe en el ser humano, más allá de las épocas, el deseo primordial de ser reconocido, porque sobre la base del reconocimiento es que se edifica toda nuestra subjetividad.

Una suma de identificaciones diversas servirá como base para la estructuración en el sujeto de la instancia psíquica llamada “Yo”, identificaciones basadas en imágenes virtuales creadas por cada sujeto. Es su reflejo, su imagen anticipada, lo que constituirá su ser…. Identificación que permitirá a un sujeto reconocerse siendo él mismo en la imagen especular . La investidura en la imagen virtual es el tiempo fundamental de la relación imaginaria, identificación narcisista que se plasma en registro simbólico en la cuestión que el sujeto dirige al Otro en término de pregunta: ¿ cómo quieres que sea?...¿Qué objeto debo ser para el Otro?....El deseo de un sujeto se instituye como deseo del Otro. El deseo se constituye entonces como el resto que queda de la tramitación de la necesidad por la demanda.

Es entonces en base a este proceso psíquico profundo, inconsciente, que cada quién asume su propia identidad. Proceso psíquico que se solidifica en la temprana juventud acompañando los cambios biológicos que metamorfosean la propia imagen infantil.

Se deduce entonces la importancia de la imagen en la estructuración psíquica de un sujeto. En la contemporaneidad, la imagen parece que viene a colmar un vacío existencial… ¿Será la época de la pérdida de sentido que provoca un vacío que se intenta llenar con imágenes?

Dos consecuencias parecen esenciales en relación a la existencia de la virtualidad. Para delimitarlas podemos utilizar la topología del nudo borromeo, es decir, del anudamiento de los tres registros: imaginario, simbólico y real, donde lo real en tanto indecible es el lugar del sujeto, lo imaginario el lugar del cuerpo y lo simbólico el lugar del discurso. Pero la existencia de la virtualidad nos remiten sobre todo al

desanudamiento del nudo. Prevalencia de lo imaginario sobre lo simbólico con apartamiento de lo real: aún cuando sabemos que la imagen es la representación de un objeto real, hoy podemos constatar que esa relación se ha invertido y que es el objeto el que se convierte en la representación de la imagen, que ordena y hasta instaura la realidad. Más aún la constituye a tal punto que la imagen puede trasformarse en la única realidad. Así la existencia de un sujeto puede quedar supeditada a las imágenes que de sí mismo suba al Facebook…."si no estás en Facebook, no existís."

Sher [1] afirma que "A certain amount of public approval confers a boost in self-esteem and personal validation. 'Millennials are more visible on the Web,' reports Douglas MacMillan of Businessweek. 'Respondents aged 18 to 29 were the most likely to say they'd posted photos of themselves and other personal data for others to see on such Web sites as Facebook and MySpace'. Many Millennials live a significant part of their lives online, and to ask them to relinquish that way of life out of concern for their privacy would not only be unthinkable, but also detrimental to their concept of identity."

Por otra parte, la prevalencia del registro real sobre el registro imaginario, quedando apartado el registro simbólico, produce cierta imposibilidad de cuestionamiento y dialectización. Ninguna pregunta puede plantearse, sólo gobierna la vida pulsional: mirar y ser mirado, ver y ser visto.

El desvanecimiento de la ley simbólica deja a los jóvenes sin brújula, viviendo como si no tuvieran que acomodarse a un Otro, como si este no existier. Lo simbólico está en crisis debilitando la ley…consecuencia de ello los excesos propios de la época, borramiento de un orden simbólico que lleva como marca también cierta indiferencia hacia la privacidad…

En este contexto, las múltiples,abigarradas e "imaginarias amistades" del Face se convierten, intentando colmar un vacío, en un nuevo e insospechado sentido que fuerza las fronteras poniendo en riesgo, acaso, valores inherentes a la dignidad humana como lo es la privacidad.

Aún suponiendo que estas amistades conservan lo esencial que marca Aristóteles de núcleo significativo: el querer, el lazo afectivo, eso se colorea de la modalidad de la época en que las nuevas tecnologías han modificado sustancialmente las nociones de espacio y tiempo: las demandas de amor y atención a los "amigos" se tienen con el imperativo de la inmediatez. Podemos decir que el chat, el mensajito, el facebook, reproducen en la época la condición de la carta de amor. Por supuesto que con la liquidez del amor contemporáneo, rasgo del amor contemporáneo que consuena con la liquidez, es la rapidez con la que se va.

Lazos sociales que privilegian la cantidad, como bien lo explicita Sher -" They do not wish to restrict their social interactions to people in close proximity".- , muchas veces en desmedro de la profundidad del vínculo; provocando conductas exhibicionistas y adictivas. " As one Millennial blogger remarks, "getting a reaction from the masses – instead of from your couch buddy – can be more addictive than settling down for a real conversation with only one person" . In this respect, Millennials could be considered a socially exhibitionistic generation."

Atentos a la importancia que encierra el reconocimiento del otro en la dinámica subjetiva y pretendiendo hacer de este trabajo un diálogo fructífero entre, posiblemente, dos generaciones, apelamos a la conciencia generacional como

herramienta potente para convertir las diferencias entre generaciones en la base misma del propio reconocimiento.

7 Conclusiones

En su artículo, Sher plantea un desafío a las generaciones mayores afirmando, "Millennials are involved with modern internet technology in a profound way. Their behaviors are, in a sense, the first significant result of the internet's social effects. Consequently, older generations fear that Millennials' acceptance of online connectivity might lead to the end of online privacy, or more generally, a redefinition of privacy itself." Más adelante agrega, "While their mentality may seem reckless from an outside perspective, Millennials have a different way of seeing the privacy issue. … It is precisely because Millennials rely so heavily on social media that many are so self-conscious regarding the information displayed about themselves. To preserve the integrity of their online identities, Millennials take advantage of built-in privacy controls regulate the visibility of their information." [1]

Si bien reconocemos la precisión de sus observaciones, sin embargo estamos preocupados por la tendencia de reducir las cuestiones relativas a la privacidad a un simple asunto de la regulación de comercio. En la medida en que las actitudes caracterizadas por Sher consiguen amplia aceptación, tememos que la hipótesis operativa será que son actitudes compartidos por todos, algo que puede tener consecuencias profundamente serias si se conduce a la revelación irreflexiva de información sensible acerca de un amigo o conocido. Estamos igualmente preocupados por la tendencia a devaluar el concepto de amistad ideal o perfecta en términos aristotélicos. Aunque reconocemos el gran valor de la posibilidad de conectar a las personas con una relación real que viven distantes uno del otro, aquí, también, la dinámica de las redes sociales parece presentar el peligro de reducir esta profunda relación a algo más parecido al comercio de información.

Bibliografía

1. Sher, Matthew, Millennial Dissonance: An Analysis of the Privacy Generational Gap. *Proceeding de II ETHICOMP Latinoamérica, CACIC.2012*, Bahía Blanca, Argentina
2. Wittes, Benjamin, Databuse: Digital Privacy and the Mosaic, accesible en línea en http://www.brookings.edu/research/papers/2011/04/01-databuse-wittes/ (2011)
3. Gavison, Ruth, Privacy and the Limits of the Law, " Yale Law Journal, vol. 89 (1979-1980), pp. 421-471.
4. Bloustein, Edward J., "Privacy as an Aspect of Human Dignity," in *Philosophical Dimensions of Privacy: An Anthology*, edited by Ferdinand Schoeman, Cambridge University Press, (1984), pp. 156-202
5. Aristóteles: *Etica Nicomaquea*. Ed. Porrúa. México, (1999)
6. Lacan, Jacques: *Escritos I*, Siglo XXI Ed. México, (1992)
7. Ministerio de Educación de la Nación Argentina. Los adolescentes y las redes sociales. Septiembre 2010

Complexity is Free, but at What Cost?
A Survey of the Current Uses of 3D Printers and the Ethical Concerns that Will Arise from Their Continued Use

Kelly Gremban
Department of Computing Sciences
Villanova University
Villanova, Pennsylvania 19085, U. S. A.
kgremb01@villanova.edu

Abstract. As 3D printing becomes more widespread, ethical decisions must be made in regards to how the technology should be used. I discuss various ways 3D printers are currently being used, and how they may be used in the future. Three ethical concerns are addressed: 1) intellectual property rights, 2) the printing of plastic firearms, and 3) the printing of living body parts.

Keywords: 3D printing, additive manufacturing, bioprinting, printable guns

1 Introduction

3D printers have been around for nearly three decades, but they are mostly used for commercial manufacturing and were made available to consumers only in recent years. As the technology becomes more versatile and affordable, it is increasingly apparent that 3D printing will be the next invention to revolutionize societies and economies worldwide. 3D printers can create everything from customizable prosthetic limbs that fit better than generic models for a fraction of the cost, to lifelike action figures, to replacement parts for out of production items. Theorists predict that they could "revamp the economics of manufacturing and revive … industry as creativity and ingenuity replace labor costs as the main concern around a variety of goods" [1]. But with technology that promises more uses than can even be comprehended at this point, there are a lot of questions to be answered, such as whether 3D printers will positively or negatively affect society, and what limitations will or should be placed on their use.

This paper will first examine the promises of 3D printing technology to revolutionize the manufacturing industry and economy, and then address two ethical concerns that will come up as this technology advances: intellectual property infringement and the use of 3D printers to create ethically debatable items, such as body parts and firearms.

2 A Brief Background

3D printing is revolutionary because it combines computer-generated ideas with effective and easy manufacturing to create products previously thought impossible. To create with a 3D printer, the user starts with a Computer Aided Design (CAD) which is a digital model of the object. These can be made by creating the design through a CAD program or by using a 3D scanner to create a model of a real-life object [2]. The CAD software then slices the model into minute cross-sections that are fractions of a millimeter thick. The printer takes these cross-sections and applies them through a process called Additive Manufacturing (AM) in which each layer of material is deposited and fused with the layer below it [3]. Printers currently on the market are mostly restricted to printing with plastics, although larger industrial printers can work with metals. According to Hod Lipson of Cornell University, "any material you can squeeze, melt or generate into a powder, you can print" [4]. There are numerous unique variations on the typical plastic or metal printers:

- The "candyfab" uses granulated sugar to print candy [5].
- The "Burritob0t" can print customized burritos in less than five minutes [6].
- The "D-Shape" prints sandstone to create houses [7].
- Researchers have printers that use living cells to print "cartilage, meniscus of the knee … spinal disks and heart valves" [4].

One benefit to 3D printing is that it is more eco-friendly than traditional methods of manufacturing. AM is revolutionarily efficient, both in terms of environmental impact and production cost. First, AM requires as little as one-tenth the amount of material as conventional approaches. Whereas traditional manufacturing must remove excess, AM builds up materials until it forms a whole [8]. Second, taking the manufacturing out of the factory also means that objects can be created anywhere, thereby cutting down on shipping requirements. Third, producing only when required removes the need for an economic system based in mass production that leads to thousands of surplus products being wasted [4]. 3D printing is beneficial because it allows for manufacturing physical objects on-site with minimal waste.

The second benefit of 3D printing is cost-efficiency for individual businesses because it streamlines the production process. Wohlers Associates, a consulting company that pays special attention to 3D printers, estimates that businesses using these devices can reduce costs by 50% and time requirements by nearly 70% [9]. The driving factor behind the cost reduction is that "complexity is free" [4]. It used to be that fabricating businesses spent most of their time creating and re-creating prototypes, and the more complex an object the more time, personnel, and money it required. With 3D printing, the major expense for companies is now just the amount of material needed to build the object [10]. 3D printing also cuts out assembly lines because a 3D printer can print moving parts at the same time, already assembled [2]. Daniel O'Conners demonstrated this by printing "a spinning gyroscopic thingumabob complete with moving ball bearings" in one session which moved freely after being removed from the machine [11].

The ability to handle complexity leads to the third benefit: innovation. With 3D printers, manufacturers and even at-home amateurs can create structures that would be impossible with any other approach. For example, a 3D printer can create a complete

bike chain, printed with the links already connected. And with reduced barriers to participate in 3D manufacturing anyone with access to a printer can contribute [8]. This change has begun to bring about the democratization of manufacturing, which, as it continues, will "allow local entrepreneurs to solve all kinds of problems, both big and small" [12]. People will not have to rely on off-the-shelf products but will be able to customize existing items or create entirely new products to find more efficient solutions. The most important function of 3D printers is the fact that they allow an entirely fresh generation of ideas to come into being. In Lipson's words, "it's not about how you duplicate things that you make today with other techniques, but it's how you explore, as we said, the new frontiers of design, making things you can't imagine today" [4]. 3D printers are not simply going to change how products are made, but will widen the definition of what it is possible to make.

As with the arrival of any revolutionary technology, the changes that come about may be difficult to embrace at first. 3D printers will change the way we think about modern manufacturing practices, and as a result could render many of them obsolete. Businesses that rely on the current way of doing things – like assembly lines and mass production – may have a hard time keeping up, but eventually this revolution will bring about new opportunities. "As businesses, industries, and jobs go away, new ones appear, and historically the new ones more than make up for the old ones that have vanished" [5]. One possible outcome is the strengthening of small businesses. Currently, it is difficult for locally-owned shops to compete with mega-store corporations. Small businesses cannot stock the same variety of products or rely on a national or global infrastructure to get cheaply produced goods. But with 3D printing, creativity will quickly surpass mass productivity in economic importance. In terms of the effects 3D printers will have on businesses and the economy, the outlook is positive.

3 Intellectual Property Concerns

Because 3D printers are so efficient at production and reproduction, there are several ethical concerns that must be addressed in the upcoming years. The first is that of intellectual property. Like the printing press, photocopier, VCR, and DVR before it, the 3D printer will be the center of a debate between individuals protecting fair use and open sources, and companies protecting copyright and patents.

The 3D printer of today is comparable to the computer in its formative years: this technology has the potential to revolutionize the creation and distribution of physical objects just as computers revolutionized the creation and communication of ideas. However, the same pitfalls that the computer industry went through have the potential to affect the 3D printing industry before it really gets started. Michael Weinberg, an attorney for Public Knowledge, refers to laws like the Digital Millennium Copyright Act that restricted the rights of the general public on the internet before the general public even knew they had those rights. He says that unless people actively learn about and defend their rights to fair use and open source materials in regards to 3D printing, they may lose them as corporations and industries that feel threatened by innovative technology try to protect themselves by restricting usage. [2]

Just as with the computer industry, the rise of the 3D printer will most likely expand the manufacturing industry but there will be strife before this can happen because it goes against most of the prevalent business models. Entrepreneurs and hobbyists looking to make use of 3D printers will have to compete with established industries protecting their interests. Patent holders will try to put restrictions on CAD files to prevent users from either scanning and reproducing copyrighted products, or creating products that infringe upon established patents. Currently, there are multiple sites where users can freely share CAD files in peer-to-peer communities. If the files become legally restricted then these open source communities may be destroyed by those who assume that any CADs shared are pirated, in the same way that Napster and other peer-to-peer sites were taken down.

In his essay, "It Will Be Awesome If They Don't Screw It Up", Weinberg advises 3D manufacturers on how to practice their rights without infringing on copyrights, patents, or trademarks. The best way to keep 3D printing technology from being restricted is by knowing how to use it without violating intellectual property rights in the first place. However, it is still going to be difficult to maintain the right to freely create and share in the face of large industries that feel threatened. The fact that there is no way to prove the benefits 3D printing will have does not make this problem any easier, because "policymakers and judges will be asked to weigh current concrete losses against future benefits that will be hard to quantify and imagine" [2]. It is likely that this case will go the way of its predecessors, photocopiers and VCRs, and be settled in favor of the new technology, but given the counter-examples of the computer industry's heightened restrictions it would be prudent to be vigilant about the public's rights to fair use of products and open source sharing of creative material.

4 Issues Arising from Printed Weapons?

While the right to creation via 3D printing should be preserved, there are some scenarios in which advanced home manufacturing could cause a real danger to the public. One benefit to the current system of centralized manufacturing is that it can be regulated. Dangerous objects like firearms are supposed to be made and distributed only by certain people and only in accordance with specific guidelines. Decentralized 3D manufacturing can avoid these regulations entirely by allowing individuals to print their own weapons, or at least enough of the component parts to avoid regulation.

In the United States, there are currently several layers of law enforcement surrounding the creation, distribution, and purchase of firearms at both the state and federal level. Specifically, "anyone 'engaged in the business' of manufacturing, importing or dealing in firearms is required to become a federal firearm licensee" and when any gun is sold, the distributor must run a background check on the buyer and record the serial number of the gun which must be included by the manufacturer. However, once you go beyond that the regulations become more complicated. For example, since the component parts of guns can be sold separately, the piece that is legally considered the "firearm" is the central frame, also known as the lower receiver, because it allows for the combination of the other pieces. Additionally, there are restrictions about how a gun may be made or what materials must be used. For

example: “the Undetectable Firearm Act of 1988 requires that all major gun components generate accurate depictions in x-ray machines and also requires assembled firearms to trigger metal detectors”. In this way, the distribution of firearms is restricted by limiting who can buy or sell guns as well as by specifying how gun parts must be made and ways to track them. [3]

The current system of regulating gun access and use is not perfect, but 3D printing is poised to upset any efficacy of the regulations. This is in part because of two current trends: 3D printing is becoming more advanced and widely available, and the firearm industry is beginning to use more polymer materials in weapons design, specifically in the design of the frame – the one regulated component. While there are still metal components that would have to be purchased from an arms manufacturer, the frame could be printed at home without having to adhere to any regulations. Additionally, 3D printers soon will have the capability of printing the highly-regulated parts that can alter a semi-automatic rifle into a fully automatic one. These abilities to make alterations at home circumvent laws restricting the use of highly dangerous weapons by the public. [3]

Americans have never been explicitly prohibited from creating their own firearms, but historically being able to make these weapons required the dedication to learn metalworking first. Now 3D printers are making it so that “a person with little to no understanding of firearms will nonetheless be capable of wielding a weapon in [a] short matter of time” [13]. This year, the Texas-based group Defense Distributed, headed by Cody Wilson, successfully fired their 3D printed handgun, “The Liberator,” and put the CAD files online for others who have 3D printers to use. The gun is entirely plastic except for a firing pin and the ammunition – the plans do include a piece of steel that would set off metal detectors, but it is an enhancement that can be omitted without affecting the functionality [14].

This 3D printing innovation has set authorities scrambling to counteract the effect of do-it-yourself, undetectable firearms. New York Congressman Steve Israel called for the renewal of the Undetectable Firearms Act after hearing the news, while New York Senator Charles Schumer suggested banning 3D-printed guns entirely. The Australian police force released a statement warning people that using the Liberator would put their personal safety at risk, explaining that they had tested it and the gun exploded on the second round. Police Commissioner Andrew Scipione attributes the “catastrophic failure” to a lack of standards for homemade weapons that endanger the gun owners as much as their targets [15]. The general political tone seems to be leaning towards restriction, but in America at least, forbidding the personal production of weapons may be constitutionally impossible.

Although there is a legitimate threat to the public involved with this system of printing, any legal action in America restricting access to or use of 3D printers may go against the Constitutional right to bear arms. In fact, there is an argument that allowing 3D printed guns would actually enrich Second Amendment protections. Currently, the right to bear arms does not apply to those who are handicapped and cannot use generically-produced weapons to defend themselves. With infinitely customizable design options, 3D printers could extend this right by creating unique guns that compensate for the user’s limited abilities [3]. Additionally, the recent Supreme Court case of the District of Columbus v. Heller upheld firearm rights, and explained that the continued right to be able to resist tyranny is one of the key reasons

for upholding the Second Amendment even in the modern age. It is arguable that the "ability to make one's own weapons, spare parts and ammunition would be essential to sustain protracted resistance against tyranny or to obtain meaningful protection in times of anarchy" [3]. If this issue goes to court in the United States, it is reasonable to expect that this argument will be made and supported by those who view their right to bear arms as inalienable.

As with all technology, there are ways to use it for dangerous purposes, but that must be weighed against the improvements and greater rights that it provides as well. That being said, protecting lives should be held above protecting rights. Until firearm regulations and police enforcement are prepared to handle the possibility of homemade weaponry, these uses for 3D printers should be pursued carefully.

5 Printing the Biological World

Printed weapons are a concern that is being addressed currently, but there are benefits to looking ahead and giving consideration to applications of 3D printing technology that are not yet affecting mainstream culture. Researchers in the medical field are using printers in ways that will revolutionize health and wellness. So far, results are still experimental, but intentions and predictions for where this technology will go next range from improving the quality of life to altering the construction of the human body.

The 3D printing of biological material, or bioprinting, uses live cells and specially designed cultures as the "ink" in their printers. This field of study shares many of the same principles as AM, but has several differences and difficulties that come about from using living material. The first is that the cells settle and readjust after being printed. For this reason, a CAD made from a scanned organ cannot be printed as-is; "the organ blueprint must be larger and probably have a slightly different shape" due to "postprinting remodeling associated with tissue fusion, tissue compaction and tissue maturation processes" [16]. The second major difference is having to prevent damage from happening to the cells during and after the printing process. Vladimir Mironov, the director of the Advanced Tissue Biofabrication Center at the Medical University of South Carolina, explains, "[f]rom an engineering point of view, high temperature and toxicity (typical for rapid prototyping technologies and processes) are not acceptable for the bioprinting process" [17]. Every step of the process of printing puts strain on the cells, from being stored in cartridges, to being ejected, to surviving in lab conditions afterwards.

Despite these complications, scientists have already had success with their bioprinting experiments:

- Laurence Bonasar of Cornell University used a modified Fab@Home printer to print cartilage directly onto a bone [17].
- A team of researchers, also using a Fab@Home 3D printer, used cartilage from calves and silver wire to print a pair of functioning bionic ears which continued to perform for more than ten weeks [18].
- The University of Bordeaux was the first to work on printing bone tissue [17].

- A group from the Wake Forest Institute for Regenerative Medicine successfully printed skin onto live animals and showed that the procedure cut the healing time of wounds by more than half [17].
- The research company Organovo printed a functioning, miniature human liver using a proprietary 3D printer, NovoGen [19].

Much of this bioprinting is focused on one of two goals: printing entire organs for transplant or printing functional tissue for medical research. Achieving the goal of printed organs will be very difficult, but steps are already being made. For example, Mironov and his co-authors state that the most challenging step is managing to print the system of arteries necessary for maintaining cell life [16], but Mironov himself goes on to claim in a later paper that several universities, including his own, have data to show that this is feasible [17]. Eventually scientists want to reach the point where they can collect a patient's cells and print a new organ directly into the body, which would have numerous benefits. Most notably, it would "once and forever eliminate patient waiting lists for organ transplantation," thus saving countless lives which would otherwise be lost simply due to lack of resources [16]. Additionally, being able to collect and print with the patient's own cells would eliminate the dangers of the body rejecting the new organ or developing tumors [17]. This achievement will allow for a higher quality of life for a greater number of people, without requiring sacrifice or endangering the patient unnecessarily.

Bioprinting tissue for medical research is likely to be happening sooner than made-to-order organs. It may not be as accurate as in-depth clinical trials with real patients, but it is expected to be "more predictable than small or even large animal testing" and simultaneously "reduce the costs of drug development and improve drug safety" [16, 17]. Overall, this will be a benefit to the medical community. Researchers will be able to have similarly if not more useful information from testing with human tissue, all without the ethical conundrum of weighing benefits against testing on sentient animals.

Since bioprinting is still in its formative years, there is a great deal of thought being put towards how it will be used in the future, and how it may even have an influential role in the shaping of the future. Mironov speculates that being able to make body parts to order with your own cells will lead to two outcomes: on one hand it can extend the length of human life as each part that wears out is replaced, and on the other hand it may create a culture of what he terms "body fashion" as people with the means to do so design and print custom body enhancements for themselves [17]. This one example demonstrates how a single piece of technology can have such far-reaching effects as to influence both the quality of life and changes in culture.

Along with expanding the length of our lives, bioprinting and 3D printing can help to expand the physical capabilities of humans. This invention could be what makes long-term space exploration possible: the ability to travel with a full hospital and manufacturing facility. If that does not sound enough like science fiction, there are some people who are interested in bioprinting for even more futuristic reasons. The group that created the bionic ears out of cartilage has explained that their goal is to develop "a unique way of attaining a seamless integration of electronics with tissues to generate 'off-the-shelf' cyborg organs" [18]. These possibilities are even spreading into the art world. Heather Dewey-Hagbog has created a work called "Stranger Visions" in which she collects discarded DNA from public places, analyzes the

samples, and then uses the genetic information to 3D print a face [20]. While these are not totally accurate resemblances – no one has recognized themselves in her work yet, at least – and she only prints in plastic, she believes that this is just a precursor to being able to clone a person from a bit of hair or skin. These predictions may seem like something from a strange tale, but researchers are working every day on turning them into reality.

6 Conclusion

3D printing is the next pivotal step in advancing technology, and will disrupt the established systems as thoroughly as the computer and the invention of the Internet did mere decades ago. 3D printing has the potential to greatly improve our quality of life and expand our ability to practice our inalienable rights. But, at the same time, there are dangers and new methods of misuse that must be anticipated and prevented. Despite the pitfalls that arise with any new piece of technology, 3D printers will benefit the quality of life. While precautions must be made to prevent dangerous and illegal use, they should not include restricting the distribution and creative freedom that will bring about innovative advancement. As enthusiasts start experimenting with strange and exciting new uses "the best improvements will spread fastest, in a process akin to Darwinian natural selection" [5]. For that reason, it is important to encourage a "diversity of approaches and strong competition among different approaches" in order to ensure superior results going forward [16]. The next several years will be crucial in the formation of 3D printing rights and restrictions, and hopefully lawmakers, industry leaders, and everyday users will work to create the most creatively supportive community possible and allow the new possibilities it opens up to develop.

Acknowledgments

Many thanks to Dr. William Fleischman, who encouraged me to submit this paper and offered support and advice for its development.

References

1. Vance, Ashlee. "3-D Printing Spurs a Manufacturing Revolution." *The New York Times* 13 Sept. 2010: n. pag. Web. 10 Apr. 2013.
2. Weinberg, Michael. *It Will Be Awesome If They Don't Screw It Up. Www.publicknowledge.org*. Public Knowledge, Nov. 2010. Web. 12 Apr. 2013.
3. Jensen-Haxel, Peter. "3D Printers, Obsolete Firearm Supply Controls, and the Right to Build Self-Defense Weapons under Heller." *Golden Gate University Law Review* (2012): n. pag. *LexisNexis*. Web. 10 Apr. 2013.

4. Wohlers, Terry, Bre Pettis, and Hod Lipson. "Can 3D Printers Reshape the World?" Interview by Ira Flatow. *Science Friday*. NPR. 22 June 2012. Radio. Transcript.
5. Easton, Thomas A. "The 3D Trainwreck: How 3D Printing Will Shake Up Manufacturing." *Analog Science Fiction & Fact*.Nov. 2008. Web. 20 Jun. 2013.
6. Cheshire, Tom. "BurritoB0t: the 3D printer that creates Mexican snacks in five minutes." Wired.co.uk. 16 Aug. 2012. Web. 18 Jun. 2013.
7. Steadman, Ian. "The race to build the first 3D-printed building." Wired.co.uk. 4 Jun. 2013. Web. 21 Jun. 2013.
8. "Print Me a Stradivarius." *The Economist* 10 Feb. 2011: n. pag. *The Economist*. Web. 15 Apr. 2013.
9. Quittner, Jeremy. "How 3D Printing Is Saving This Jewelry Design Business." *Crain's New York Business*. N.p., 20 Oct. 2010. Web. 15 Apr. 2013.
10. Graham-Rowe, Duncan. "3-D Printing for the Masses." *MIT Technology Review* (n.d.): n. pag. 31 July 2008. Web. 12 Apr. 2013.
11. O'Connor, Daniel. "Thingi Thursday: Spinning Gyroscope." Weblog post. *Www.prsnlz.me*. N.p., 13 Jun. 2013. Web. 16 Jun. 2013. http://www.prsnlz.me/blogs/daniel-oconnors-blog/thingi-thursday-spinning-gyroscope/.
12. MacDonald, Chris. "3D Printing and the Ethics of Value Creation." Web log post. *The Business Ethics Blog*. N.p., 1 Dec. 2012. Web. 10 Apr. 2013.
13. O'Neill, Kevin J. *Is Technology Outmoding Traditional Firearms Regulation? 3-D Printing, State Security, and the Need for Regulatory Foresight in Gun Policy*. *Social Science Research Network*. N.p., 3 May 2012. Web. 16 Apr. 2013. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2186936>.
14. Greenberg, Andy. "Meet The 'Liberator': Test-Firing The World's First Fully 3D-Printed Gun." *Forbes*. N.p., 5 May 2013. Web. 28 June 2013.
15. Clark, Liat. "Australian Police: Exploding 3D Printed Gun Will Kill You and Your Victim." *Wired.co.uk*. Wired, 24 May 2013. Web. 25 June 2013.
16. Mironov, Vladimir, Vladimir Kasyanov, Christopher Drake, and Roger R. Markwald. "Organ Printing: Promises and Challenges." *Regenerative Medicine* 3.1 (2008): 93-103. Web. 25 Jun 2013.
17. Mironov, Vladimir. "The Future of Medicine: Are Custom-Printed Organs on the Horizon?" *The Futurist* 45.1 (2011): 21-24. *Proquest*. Web. 25 June 2013.
18. Mannoor, Manu S., Ziwen Jiang, Teena James, Yong Lin Kong, Karen A. Malatesta, Winston O. Soboveio, Naveen Verma, David H. Gracias, and Michael C. McAlpine. "3D Printed Bionic Ears." *Nano Letters* (2013): 2634-639. *Pubs.acs.org*. 1 May 2013. Web. 27 June 2013.
19. Organovo. *Organovo Describes First Fully Cellular 3D Bioprinted Liver Tissue*. *Ir.organovo.com*. N.p., 22 Apr. 2013. Web. 27 June 2013.
20. Dewey-Hagborg, Heather. *Stranger Vision*. N.p., 7 Mar. 2013. Web. 24 June 2013. <http://deweyhagborg.com/strangervisions>.

Greedy for Green: The Motivations and Hindrances of Maintaining Environmental Sustainability in the Commericial Industry of Cloud Computing

Bradley S. Cantor
Enrolled in the Master of Software Engineering Program
Department of Computer Science
Villanova University
Villanova, Pennsylvania 19085, U. S. A.
bcanto01@villanova.edu

Abstract. This articleexploresthe potential for environmental sustainability in the rapidly growing industry of Could Computing. By understanding the Cloud as a commercially driven endeavor, we unearth the motivations that drive the growth of the burgeoning industry. We go on to explain the impetus behind many of the Data Center innovations that Cloud Computing has brought about. We show that these innovations are why the Cloud has been heralded as a form of Sustainable Computing.After exploring the technologies endorsed and innovated by Cloud Computing, the article proceeds todiscuss some of the potential downfalls inherent to propagating a commercially driven, yet environmentally sustainable industry. Lastly, the paper discusses the role that we, the consumer, play in shaping the environmental sustainability of the digital storage industry.

Keywords: Cloud Computing, Sustainable Computing, Data Center innovations, ethical consumerism, economic viability

1 Introduction

We live in an era of incredible innovation. The speed of this innovation is almost overwhelming. There are few among us who have not felt the remorse brought about from buying a brand new smartphone or flat-screen TV only to learn days later that a sleeker, fancier model has been announced. This type of rampant technological

growth defines the past few decades. Landlines have given way to cellphones. Dial-up modems are all but extinct. Wi-Fi flows wherever there is coffee (just be careful asking for 'Java'). High-schoolers can become 'friends' with a single click and a series of hash-tags is currently creating the next big, yet brief, internet trend. The speed of change is hard to keep up with and impossible to ignore. #innovation

Constantly evolving gadgets, software applications and websites have become an integral part of daily life. This has created a society in which technology is no longer optional. The mail cluttered kitchen counters of our parent's generation have been replaced by our own poorly managed and often overflowing digital inboxes. Our computers act as our rolodexes, our file cabinets, our calendars, book shelves and CD collections. The stuff that used to be jumbled around our homes now fills our hard drives. The movement to the digital environment has altered the way that we think about storing information. Fifteen years ago, nobody would have concerned themselves with bringing two thousand CD's and fifty seven books on a two hour train ride. The advent of smart phones and tablets has changed our perception of accessibility. We want everything we own to be everywhere we go, and we want to be able to access it instantly. The rigid upload/download format of yesteryears MP3 players has given way to a far more dynamic interface of personal data storage; the cloud.

Entire archives of data, artwork, and entertainment flow through a network of revolving doors that connect millions of computers, phones and handheld devices around the world. The sheer magnitude of data has turned 'space' into a digital commodity. It is sometimes easy to forget that this data does not just magically float around in the sky. The premium placed on space has created an entirely new storage industry. Data centers, or server farms, have begun to pop up all over the United States, Europe and Asia. These data centers power the cloud. Without them, there would be no Spotify, no Amazon Web Services, or Google App Engine. Every megabyte of information 'floating' in the Cloud is stored in one of these servers [1]. This means that as the Cloud continues to grow, more and more data centers will be designed, built, and utilized. This growth represents incredible commercial opportunity. We are already beginning to see a mad dash for tech companies to provide Cloud services. But the Cloud model embodies more than just economic potential. This new industry could very well help usher in a new age of corporate responsibility and environmental sustainability.

The way that data centers are designed, built, and utilized will determine both the economic and environmental potential of Cloud Computing. As the industry grows it has the capability to redefine standard best practices in the field of Sustainable Computing. Energy efficient data center design and innovative server technologies have seemed to evolve alongside the rapidly growing Cloud but these practices are in no way guaranteed. As with all commercial endeavors, profitability will inform practice. In the rampantly growing industry of digital storage, economic viability will supersede environmental ideology. The potential for the Cloud to usher in a new era of Sustainable Computing will rely not on ethical responsibility, but instead on dollars and cents. The question is not whether the ethicality of the Cloud can stand up to corporate consumerism but instead whether the two coincide.

2Sustainable Computing

The field of Sustainable or 'Green' computing is "the study and practice of designing, manufacturing, using, and disposing of computers, servers, and associated subsystems—such as monitors, printers, storage devices, and networking and communications systems — efficiently and effectively with minimal or no impact on the environment[3]". While this definition sums up many of the goals of Sustainable Computing there is another side of the environmental endeavor that often gets overlooked. Sam Murugesan, the author that provided the aforementioned definition, writes that "Green IT also strives to achieve economic viability and improved system performance and use, while abiding by our social and ethical responsibilities [3]".This aspect of Sustainable Computing perfectly depicts what it is that makes the Cloud so interesting.The future of Sustainable Computing depends on the potential of economic viability. The mere practice of designing, using and disposing of hardware in an environmentally friendly way will not, in itself, provide an incentive for large-scale change. The same can be said for Data centers. If the green option is more expensive or less effective it will be unable to compete in the market. Cut-throat corporations are not going to trade economic viability for environmental sustainability. While this may seem problematic for green computing and the future of the Cloud, economic viability and environmental sustainability are not mutually exclusive.

3Incentives for Cloud Computing

A survey conducted by Sun Microsystems Australia involving 1,500 responses from 758 large and small organizations in Australia and New Zealand, found that reducing power consumption and lowering costs were the major reasons that companies chose to incorporate environmentally responsible practices. The environmental impact of these practices was something of an afterthought [4]. Businesses prioritize environmental sustainability if and when it is cost effective. When corporate concerns (cost and profit) align with sustainable practices the motivation for these practices need not matter. Sustainable Computing on the corporate level does not need to be a morality contest. Corporations will choose eco-friendly practices when they are the most cost effective. That is the bottom line. The environmental benefits are merely a byproduct.

This is where the innovations of Cloud Computing come into play. The Cloud presents a potential for economic viability and Sustainable Computing to coincide. In the past, corporations, both large and small, relied on small on-site data centers. That meant that in some form or another, companies were forced to use office space to physically house their servers. The databases, websites, and software that allowed businesses to function were all stored in these localized servers. Any loss of server functionality was potentially devastating; an onsite crash could instantly debilitate a business. This meant that corporations had no choice but to heavily invest in the maintenance and upkeep of their onsite data centers. Every localized data center needed its own team of IT professionals to provide technical support for the servers. In addition to these IT professionals, these small in-house data centers required

housing specifications and environmental control systems that far exceeded the ordinary demands of generic office space [5]. Insulation and temperature control was vital to preserving optimal functionality. The introduction of Cloud Computing mitigates most, if not all, of these costs.Companies can spend less on IT professionals, optimize office space and, most of all, avoid incredibly inefficient energy expenditures.

By outsourcing digital storage, companies are able to streamline costs.According to a recent research report by Accenture, small businesses experienced a reduction in emissions of up to 90 percent while using Cloud resources. Additionally, large corporations also saw improvements of 30-60 percent in carbon emissions while using Cloud applications[6]. But why is the Cloud environmentally beneficial? It would seem that it just moves energy expenditure from one place to another. The Cloud model simply replaces two million, small-business onsite data centers with a few massive server farms. Perhaps there are fewer data centers but the amount of information that has to be stored is still the same. This would seem to suggest that the same amount of servers, and therefore, the same amount of net energy would still be required. After all, the sum of the parts is always equal to the whole. While it is hard to say that this line of thought is not highly logical, it is also flawed. The truth is that Cloud Computing represents far more than just a geographical relocation of data centers. Technological innovations brought about by the Cloud are changing the way that servers are stored and utilized. It is these innovations that are minimizing energy expenditure and therefore aligning corporate incentive with environmental ethicality.

4Innovations in Cloud Based Data Centers

Emerging standard best practices are enabling Cloud based data centers to become increasingly energy efficient. Pre-Cloud on-site data centers were designed to handle sporadic peak loads. The problem with this approach is that during non-peak hours there was no way to reduce operational functionality. This reduced resource utilization resulted in wasted energy. Data centers that use emerging cloud practices, on the other hand, can greatly increase resource allocation through server consolidation. Different companies can share the same server using a parallel processing and partitioning method called virtualization. This practice helps alleviate energy consumption that would have otherwise been spent on the electrical costs of running a local server during non-peak hours. As workloads are consolidated onto partitioned servers, unused servers can be switched off. Before virtualization, this approach was impossible. Each server in an onsite data center had its own data, its own files and therefore its own function. It is only the dynamic partitioning of virtualization that allows for the ebb and flow of server consolidation to conserve energy [7].

The innovations brought about by Computer Science are in themselves remarkable, but the Cloud also presents a potential for sustainability that is far more intuitive. Other than the power required to physically run the servers, the largest energy expenditure experienced by data centers is the cost of temperature control [8].

Before the Cloud, data centers were generally a converted room or collection of rooms in an office building. The center was almost certainly retrofitted to house servers; it was not the original purpose of the structure. The state-of-the-art server farms that are now being utilized by companies that provide Cloud services are cut from a very different cloth. Buildings are being designed for the sole purpose of storing servers. The energy costs of cooling the servers can be alleviated by choosing sites that remain cold throughout the year. Harnessing wind, shade, water, and ground temperature can help lessen operational costs. Due to the energy demands inherent to data centers, infrastructure siting is becoming increasingly dictated by climate and resource allocation. Investing in these state-of-the-art centers is not cheap but companies that provide Cloud services seem to be increasingly willing to take this approach; spend now to save later [9].

5 Concerns for Future Sustainability

Data-centers are the infrastructure of Cloud Computing. They play an integral and inevitable part in the business of digital storage. By taking climate and ecological factors into account, data centers are able to reach new levels of energy efficiency. This green infrastructure is not only cost effective, but it also provides a strong anchor for the future of Sustainable Computing. After all, these new data centers are the roots on which the Cloud must grow. But a recent Greenpeace report released in April of 2012 is far more skeptical about the inherent greenness of "the factories of the 21st century information age. [9]" Weary of the potential downfalls of massive data centers and the large corporations that own them, the Greenpeace report stresses the importance of energy transparency, infrastructure siting, and the use of clean/renewable energy. While the findings do suggest a trend towards corporate environmental advocacy, the report also argues that the Cloud phenomenon is not necessarily, in and of itself, eco-friendly.

It should come as no surprise that the issue is energy. The electrical cost of maintaining the aggregated digital information at present is already beginning to become astronomical. The expected growth of Cloud Computing over the coming years will dramatically increase this cost. According to the Environmental Protection Agency, "data centers now account for 1.5 percent of all electricity consumption in the U.S. and by 2020, carbon emissions will have quadrupled to 680 million tons per year, which will account for more than the aviation industry [10]". Furthermore, a recent estimate of the IT sector's footprint conducted as part of the 2008 SMART study concluded that data centers will be responsible for 2% of global GHG emissions by 2020 [11]. Incredibly, the Greenpeace report also estimates that the industry will experience "a 50-fold increase in the amount of digital information by 2020 and nearly half a trillion in investments in the coming year [11]".

These numbers are staggering. They illustrate the profound technological and commercial role that Cloud computing will play in the coming years. In the face of such rapid expansion, Greenpeace, and many others, worry that Cloud providers will become increasingly fixated on lowering the operational energy costs of their data centers. There is no question that infrastructure siting and design innovations will help

accomplish this but in business, profit informs growth. State-of-the-art data centers are expensive and take years to begin to pay for themselves. In the mad dash to take up Cloud market shares there is a strong possibility that rising energy costs will be mitigated by purchasing cheaper fuel. Infrastructure siting is not just about harnessing climatic features; it is heavily influenced by geographical resources. A company that decides to build a new data center in West Virginia is doing so because of the price of coal, not because of the cool mountain air.

If coal becomes the main fuel source of data centers, the environmentally friendly practices innovated by Cloud technologies will be vastly overshadowed by the sheer quantity of greenhouse gases produced by the industry. Energy efficient techniques like workload consolidation and virtualization will not be enough to level out the environmental impact [12]. Fuel is the one fundamental issue in Cloud Computing in which economic viability does not align with sustainability. If data centers turn to coal, Cloud Computing will cease to be "ethically responsible". The potential for a new age of Sustainable Computing brought about by the burgeoning Cloud will fall by the wayside. Simply put, a Cloud industry that garners the majority of its fuel from coal will be environmentally disastrous.

6 Consumer Ethicality

With the continual growth of the internet and the increasing market share of server-dependent companies, the demand for Cloud technologies and data centers is going to increase substantially. Cloud computing is in its embryonic stages. The possibilities are in many ways limitless, but as everyday consumers become increasingly accustomed to the possibilities of the Cloud, the temptation for data centers to tap cheap energy sources will prove difficult to resist. Low energy costs have the potential to subvert high energy efficiency. Cheap fuel will equal cheap digital storage. In an industry dictated by profit, it would seem that this business model would be difficult to combat. If offered the option to pay five cents or ten cents for a gig of space which would you choose? For the past few years Greenpeace has issued environmental report cards for Cloud Service Providers. Companies like Amazon and Twitter have literally failed in multiple categories. Yet, these are two of the most profitable companies in the Tech industry [11][13] [14].

As the Cloud continues to grow, economic viability will determine the policies and practices ofservice provider data centers. It is easy to say that large corporations should do more to protect the environment, but what is our role in all this?At the end of the day, it is not Google or Greenpeace that will decide the future of environmental sustainability in digital storage. It is the end user. It is whether or not we care. If the environment matters to the end user it will be economically advantageous to design, build, and utilize data centers in a way that coincides with the principles of Sustainable Computing. If the end user does not care, following these principles will cease to be economically pragmatic.

Companies like Yahoo, Facebook, and Google are spending millions of dollars on new, state-of-the-art, ultra energy-efficient data centers. By tapping renewable energy sources like nuclear, wind and solar, these companies are making

strides to help ensure that the Cloud remains a green enterprise [15]. Their actions set the bar high for others to follow, but these are companies that are never too far from the public eye. Appearing environmentally trustworthy is good for business. But branding, like energy conservation, helps align economic viability with ethical and environmental responsibility.

7 Conclusions

We use the Cloud to make our lives simpler—more streamlined, organized and accessible. Cloud Computing allows us to clean off our kitchen counters, throw out our scratched old CD's, and travel anywhere in the world with a library in our pocket. But rarely do we stop and think about the cost of this convenience. Rarely do we consider the equation in its entirety. What is the cost of this digital information?

Within the Cloud, every undeleted email, every blurry family picture, every video of a cat wearing socks, is stored on a server.Innovations in server technology along with data center design and utilization have helped forge a digital storage industry that is environmentally responsible and economically profitable. The Cloud has brought about innovation in infrastructure siting, workload consolidation, and server virtualization.But, the rapid growth of the industry has the potential to undermine the environmental benefits of these innovations. The Cloud is an energy hungry enterprise. If data centers turn to coal as the main source of energy, the Cloud will cease to be 'green'.

We as consumers hold the key tothis green potential. The industry of Cloud Computingwill become whatever we make it. Our decisions, our values, our beliefs, and, most of all,our dollars, will determine the future of the Cloud. The consumer will ultimately define the practices that are profitable. If cheap storage powered by coal is demanded, the market will be sure to supply it. While this understanding of the Cloud may lack a certain ideological flare for the ethics of capital markets, it in no way condemns the Cloud to a future of bleak possibilities. We as consumers should care about the environment, and perhaps we will. But the Cloud will not stay green because it ought to, or because Sustainable Computing is ethically correct. The Cloud will stay green if and only if it is economically viable.

Acknowledgments

I would like to thank Dr. William M. Fleischman for his support and guidance throughout this process. This article originated asa term paper in the Computer Ethics class offered through the Computer Science Department at the University of Villanova. Without Dr. Fleischman's foresight and curiosity, I would never have been able to develop that paper into the article that it has now become.

References

1. Anderson, S., Improving Data Center Efficiency, Energy Engineering, Vol. 107, No. 5, pg 42-63. (2010)

2. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M., A View of Cloud Computing, Communications of the ACM, April, Vol. 53, No. 4. (2010)

3. Murugesan, S.: Harnessing Green IT: Principles and Practices, IEEE IT Professional, January–February 2008, pp 24-33.

4. Andonova, L.: Networks, Club Goods, and Partnerships for Sustainability: The Green Power Market Development Group, Colby College

5. Greenberg, S., Mills, E., Tschudi, B., Rumsey, P., and Myatt, B., Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers. ACEEE Summer Study on Energy Efficiency in Buildings. available at http://eetd.lbl.gov/emills/PUBS/PDF/ACEEE-datacenters.pdf, last accessed 10 May 2013 (2008)

6. Accenture Microsoft Report: Cloud computing and Sustainability: The Environmental Benefits of Moving to the Cloud, available at http://www.wspenvironmental.com/media/docs/newsroom/Cloud_computing_and_Sustainability_-_Whitepaper_-_Nov_2010.pdf. last accessed 21 June 2013 (2010)

7. Verma, A., Koller, R., Useche, L., and Rangaswami, R., 2010. SRCMap: energyproportional storage using dynamic consolidation. 2010, Proceedings of the 8th USENIXconference on File and storage technologies (FAST'10), San Jose, California

8. Robles, L. Four Trends that Shaped Cloud Computing in 2011, available athttp://venturebeat.com/2011/11/29/four-trends-that-shaped-cloud-computing-in-2011/last accessed 21 May 2013 (2011).

9. Greenpeace International: Make IT Green available at http://www.greenpeace.org/international/en/publications/reports/make-it-green-Cloudcomputing/, last accessed 21 July 2013 (2012)

10. Metreweli, K., Data Center Cost Savings Go Green, available at www.ebizq.net/topics/int_sbp/features/11740.html, last accessed 21 May 2013 (2009)

11. Greenpeace International: How Green is Your Cloud, available at, http://www.greenpeace.org/usa/Global/international/publications/climate/2012/iCoal/HowCleanisYourCloud.pdf last accessed 25 July 2013 (2012)

12. Garg, K., Buyya, R., Green Cloud Computing and Environmental Sustainability, The University of Melbourne, Australia

13. Forbes Magazine: Global 2000: The World's Biggest Public Companies available at http://www.forbes.com/companies/amazon/, last accessed 27 July 2013 (2012)

14. Shontell, A.: The Digital 100: The world's most valuable private tech companies, Buisness Insider 2013, available at http://www.businessinsider.com/2012-digital-100?op=1, last accessed 27 July, 2013 (2013)

15. Cook, G.,Van Horn J. How dirty is your data?, available at http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data/, last accessed 21 July 2013 (2011)

16. Comerford, T.: Principles of Data Center Siting and Economic Development Incentives, UEDA Winter Forum, Feb 23, 2011 available at http://www.blsstrategies.com/Docs/Events/Event_33.pdf, last accessed 27 July 2013 (2011)

Why We Should Not Build Autonomous Robotic Weapons

William M. Fleischman
Departments of Computing SciencesandMathematicalSciences
Villanova University
Villanova, Pennsylvania 19085, U. S. A.
william.fleischman@villanova.edu

Abstract. We discuss robotic weapons, their advantages and disadvantages, and their effect on the way humans wage war. We consider the factors favoring the development of lethal robotic weapons that can operate autonomously. We discuss the attempt to mitigate the dangers inherent in such weapons by means of an ethical controller implemented in software. We conclude that this is impossible to achieve and therefore that autonomous lethal robotic weapons should not be developed.

Keywords: Computer ethics, Robotic weapons, Autonomous moral agents.

1 Introduction

In this paper, we argue that fully autonomous robotic weapons that have the capacity to kill should not be developed or deployed. We begin with an overview of current robotic devices that are deployed or under development by the military. We discuss the advantages and disadvantages that these weapons confer in combat and in the larger context of decisions to wage war and attitudes toward the conduct of war. We consider the thorny problem of keeping humans "in the loop" in situations where these weapons are used with potentially lethal effects and discuss efforts to develop a so-called "ethical governor" to restrict the behavior of robotic weapons capable of autonomous operation. We present an instructive example from the history of the Cold War that underscores the importance of human deliberation in situations of belligerent confrontation. This example is followed by a discussion of the problem of responsibility and accountability as it applies to autonomous robotic weapons. We conclude with a section of general observations about the choices involved in opting to invest important material and human resources in the development of lethal autonomous weapons.

2 An Overview of Robotic Devices in Use and Under Development

We begin with a short discussion of the recent accelerated development of robotic weapons – unmanned ground and aerial vehicles (UGVs and UAVs) under the impetus of the wars in Iraq and Afghanistan. Numbers tell at least part of the story. There were few of either type of system deployed in the 2003 invasion of Iraq. By 2011, there were an estimated 12,000 UGVs and 7,000 UAVs in the inventory of the

U.S. military forces. Significantly, the U.S. Air Force currently trains more UAV operators than fighter and bomber pilots combined. [1]

Enemy deployment of IEDs in Iraq created an instant demand for Packbots – a ground-based, essentially defensive device developed by iRobot, the Boston-area company originally famous for manufacturing the Roomba robotic vacuum cleaner. The Packbot was used to detect and, if necessary, disarm IEDs without the risk of loss of human life. Initially, it was simply thought of as a "mobile pair of binoculars." With the addition of simple effector arms and grippers the Packbot acquired the capability to disarm and destroy improvised explosive devices concealed by the enemy. [2]

The initial problem addressed was that of locating and identifying non-human threats. A related problem, of course, is the location and identification of human threats – enemy snipers. In this application, however, once a threat is identified, the next job is to eliminate it by killing the sniper. Quite logically, a mobile device that carries a weapon in addition to its cameras provides the possibility of eliminating the threat by aiming and firing remotely under control of a soldier who does not have to appear in the sight of the sniper's weapon. Once again, the desire to shield one's soldiers from situations in which their lives are at risk provides the incentive for development of a robotic device with additional capabilities. As has often been observed [3], the desire to increase the killing effectiveness of one's soldiers while increasing the distance between them and the enemy is a constant in the history of warfare. So this was a natural application for Packbot, Warrior, its more heavily armed successor, and congeners such as the Talon and SWORDS devices manufactured by Foster-Wheeler, a second Boston-based robotics firm. In essence, arming the Packbot or similar robotic device is simply a next step in this historical process.

Not all UGVs have direct combat roles. The special dangers of the role of human medics serving in battlefield situations has led to the development of a version of the Packbot that can search for wounded soldiers and provide a video feed that allows a distant human controller to deploy medical equipment on the so-called "med-bot" in order to evaluate and treat the wounded individual. [2]

UAVs have undergone a similar rapid transformation. Perhaps the best-known UAV is the twenty-seven foot long Predator, capable of carrying out 24-hour reconnaissance and surveillance missions, returning high quality images day and night by means of normal and infrared cameras. Furthermore, the Predator's synthetic-aperture radar can provide valuable information even where the terrain is obscured by clouds, smoke, or dust. As Singer notes, [t]he exact capabilities of the system are classified, but soldiers say they can read a license plate from two miles up." [4]. The same logic that has driven changes in the design of UGVs has resulted in the arming of UAVs which now can carry out offensive missions under the direction of a human pilot or operator located thousands of miles away. In addition, UAV technology has spread in both directions along the size and mission-length continua with the Raven (thirty-eight inches in length and ninety minutes in the air), the Wasp (fifteen inches in length, forty-five minutes of endurance), with micro-UAVs the size of insects in the planning stage. At the other end of the spectrum, the newer Reaper [get some specs if you intend to include this UAV in the discussion] and the Global Hawk (nearly forty-eight feet long with an endurance of thirty-five

hours) provides both wide-area search and high-resolution single target identification and has the capability of autonomous operation between the signals to taxi, take off, and land provided by its human operator. [4]

In addition to deploying its own force of UAVs, the U.S. Navy is also developing various types of unmanned surface and underwater vessels (USVs and UUV's). [4]

3 Advantages and Disadvantages of Robotic Weaponry

It is not hard to see (and it is very hard to resist) the advantages of robotic weaponry. The first, and most compelling for an armed force possessing these weapons, is that they replace humans on the battlefield and therefore reduce the number of human casualties this force will sustain. Beyond this, they are markedly superior to humans in what military strategists describe as the "three D's" – situations that are dangerous, dirty, and dull.

Dirty environments include not only those, like desert battlefields affected by smog, smoke, sand and dust, but also those which have been contaminated by biological, chemical, or radioactive agents. Robots have a very clear advantage in these environments where humans would be encumbered by bulky protective suits and related gear.

Many military missions require concentration over long periods of time. In addition to the physical stress of the activity, there is the psychological stress of paying steady attention in otherwise boring circumstances. Humans can do this for limited periods of time and need downtime or pauses to recover the necessary level of acuity. By contrast, robots don't need to sleep, to eat, or to take a break for "rest and recreation."

The human body is limited in the speed and limits of reaction to threats and forces to which it is exposed in combat situations. From g-forces acting on human pilots of advanced aircraft to speed of recognition and reaction to battlefield dangers, robotic systems appear to have a clear advantage. As already noted, the first advantage of a robot in a dangerous environment is that its destruction involves the loss of a machine (although this may be more consequential if it falls into the hands of an enemy who can study and copy it) and not the loss of a human life.

Related to these factors is calculation regarding risk. Singer notes that, "The unmanning of [an] operation also means that the robot can take risks that a human wouldn't otherwise, risks that might mean fewer mistakes." [4] He cites friendly fire incidents during the Kosovo campaign in 1999 in which the imperative to avoid loss of NATO pilots resulted in orders that planes not be flown at altitudes below 15,000 feet. One of the most grievous errors of this nature occurred when NATO planes flying at these altitudes bombed a convoy of buses carrying Kosovar refugees mistakenly identifying them as a convoy of Serbian tanks. Singer also notes that the "removal of risk allows decisions to be made in a more deliberate manner than normally possible. Soldiers describe how one of the toughest aspects of fighting in cities is how you have to burst into a building and, in a matter of milliseconds, figure out who is an enemy and who is a civilian." In this situation, a robot that can enter a room and shoot only at someone who shoots first has a distinct advantage over the

human who must take fire and somehow instantly manage to determine the source, return fire, and avoid hitting any civilians. [4]

Another advantage that robots have in situations of combat is that they do not suffer from human emotions of rage against adversaries who have caused harm or death to a soldier's comrades. We know of many episodes where otherwise good individuals have given way to extreme emotion and committed atrocities after experiencing the loss of or grievous harm to someone with whom they have bonded and upon whom they have depended in situations of danger. Surely eliminating the danger of such episodes is an important advantage favoring robotic agents over humans.

With all these advantages noted, what could possibly be the downside of the use of robotic weapons? These may be more subtle and harder to see but, in a certain sense, the disadvantages of these weapons are identical with their advantages. One of these disadvantages, clearly recognized by those in command positions in the military, is that over a long time and haltingly we have negotiated barriers against barbaric behavior in war. The Geneva Conventions and treaties barring the use of chemical and biological weapons are among these barriers. When, however, one side in a conflict has such technological superiority, when there is marked asymmetry in the resources each brings to battle, there is an inescapable lessening of the respect that each side owes the other out of recognition of the parity of the risks the combatants share. The sense that the weaker forces can be eradicated like insects by the "magic" of advanced technology acts, in a mutually reinforcing manner, on both sides to undercut the restraints erected against barbarity. [5]

Perhaps the most serious disadvantage of robotic weapons has to do with another set of barriers. General Robert E. Lee, commander of the Confederate forces in the American Civil War of the 19th century once wrote, "It is good that we find war so horrible, or else we would become fond of it." [3] The act of declaring war is or should be a grave existential decision for any country. But we have seen, perhaps most notably in the case of the ill-considered invasion of Iraq by the United States, how consciousness of technological superiority lowers the barrier against waging war.

Paradoxically, to the extent that atrocities committed by otherwise decent soldiers of our military remind us of the horror of war, they serve as a factor that should give pause to anyone contemplating "loosing the dogs of war."

4 Keeping Humans 'In the Loop'

This section is the easiest to write and the most frightening. When it comes to giving robotic weapons lethal capabilities, official military policy seems to be very clear and emphatic: "Humans must be kept in the loop." The meaning of this is, or should be, that a human must give authorization before any robotic weapon can fire on a human target. In fact, however, whenever this matter is raised in serious discussion, the result is averted eyes and a change in topic. The reasons for this are also clear. Although the ideal is to keep humans in the [command] loop, there are so many factors militating against this that in practice it seems impractical. Why, if there is risk of loss of life on your side in the interval between identification of a lethal threat

and authorization to fire issued by a human controller, should the robotic weapon not be given the capability to fire immediately upon locating the threat? Since the authorization requires communication between controller and weapon, and this communication can be cut or disrupted by the enemy, why should there not be an emergency back-up capability for the weapon to operate autonomously in this situation?

Singer points out that the logic of human control of robotic weapons seems to demand a many-one correspondence between weapons and controllers. But humans are notoriously ill equipped and unreliable for the task of controlling multiple units at one time, even under relatively calm conditions. A Pentagon-funded report notes that, "Even if the tactical commander is aware of the location of all his units, the combat is so fluid and fast-paced that it is very difficult to control them." [3]

Further, as Singer points out, human control of automated weapons systems has already been seriously compromised by the human tendency to "believe what the computer says." The paradigmatic example of this is the case of the downing of Iran Air flight 665 over the Persian Gulf in July 1988 by an American naval vessel patrolling the gulf during the Iran-Iraq war. Iran Air Flight 665 was an Airbus passenger jet on a commercial flight from Tehran, Iran to Dubai via Bandar Abbas. On the morning of the flight, the U. S. Navy guided missile cruiser, the Vincennes, equipped with the Aegis combat system, an integrated weapon control system that uses powerful computers and radars to coordinate, track, and guide weapons to destroy enemy targets. Even though the passenger jet was climbing after takeoff from Bandar Abbas, flying a consistent course, and "squawking" the appropriate radio signal that proclaimed it to be a civilian airliner, the Aegis system radars on board the Vincennes seemed to identify the plane as an assumed enemy fighter jet on a descending attack profile. Even though most members of the crew of the Vincennes and almost everyone on board its sister ships on patrol that morning were reading data that accurately identified the nature of the flight, not one of the eighteen sailors and officers of the Vincennes were willing to question the Aegis system's apparent mistaken designation of an attacking enemy fighter aircraft. As a result, the captain of the Vincennes, an officer with a known penchant for aggressive action, gave the authorization to fire resulting in the destruction of Iran Air Flight 655, killing all 290 passengers and crew, among them sixty-six children. [3, 6]

The problem with Singer's analysis and the flaw in the conclusion drawn is that a software design or software-engineering error that should not have evaded the eye of even undergraduate software engineering students was one of the principal factors implicated in the mistaken characterization of flight 655. In fact, in the process of coordinating data on the radars of the three ships in the patrol, a tag used within the previous hour to label a (friendly) fighter jet making a landing (thus descending) was reassigned as the label for Iran Air Flight 655 on the radars of the Vincennes. [6] So while it is not entirely inaccurate to think of this as an illustration of the way in which humans defer to the "judgment" of computer-controlled systems, it is far more relevant to see this as a warning against placing too much trust in the reliability of even state of the art software engineering.

5 Compensating for the Human 'Out of the Loop'

If the superior capabilities of robotic weapons and the limitations of humans acting as controllers so far compromise the military principle of always keeping the human in the loop, then perhaps we can substitute an "ethical governor" implemented in software for the absent human controller. Properly programmed, weapons acting in autonomous mode could perhaps be constrained to "act ethically in war," observing all the articles of the Geneva Conventions, the laws of war, and, in the local context of the combat in which they are deployed, the relevant rules of engagement. And since they are not subject to the psychological and emotional stresses that affect human combatants, we might even expect that they would act more morally than the human soldiers whose combat roles they assume.

In fact, the NSF and U.S. government agencies associated with the Department of Defense have funded an initiative of precisely this nature. Singer quotes the assertion of Ronald Arkin, a professor of computer science at Georgia Tech who has received support various agencies of the government for just such a project, "Ultimately these systems could have more information to make wiser decisions than a human could make. Some robots are already stronger, faster, and smarter than humans. We want to do better than people, to ultimately save more lives." [5]

In a recent paper, Gerdes and Øhrstrom discuss the possibility of devising a Moral Turing Test, which, in their words, "might enable us to distinguish principles for evaluating morally correct *actions* rather than (as in the original Turing test) skills of articulation." [7] Such a test would constitute a necessary but not sufficient condition for the development of what is referred to as an Artificial Moral Agent. Their analysis, rooted in the work of the logician A. N. Prior [], leads to the conclusion that "Prior was right in claiming that the formulation of a formal system which correctly incorporates all aspects of moral reasoning would in principle require a complete description not only of all relevant moral rules and laws but also of all relevant aspects of the situation in question. However, having such descriptions is tantamount to having a God's eye view of all relevant aspects of reality." Although they conclude that it may still be "possible to formalize important aspects of ethical reasoning in a specific context and thereby contribute to a system which may pass a comparative Moral Turing Test," I take their paper as indicating that even this partial approach to creating an Artificial Moral Agent represents a software engineering project of considerable difficulty and complexity. Since the conditions of actual combat constitute a context of such fluidity and rapid change as to defy the simple description of "a specific [i.e., closed] context," it is not unreasonable to conclude that the project envisioned by Arkin has an impossible goal. The similarity of this case with that of the Strategic Defense Initiative (the so-called 'Star Wars' project) from which David L. Parnas withdrew in a well-known letter and series of critical papers [8] suggests that the appropriate response of computer scientists of good conscience toward Arkin's project or any other claiming to have the purpose of devising an "ethical governor" for autonomous robotic weapons should be to condemn it.

In this light, I think it is important to ask, "What is the purpose of the NSF in funding this "research?" Why should anyone want to do this? One possible motivation is as a salve to the consciences of those who are participating in and drawing public funds from the Department of Defense and the National Science

Foundation in research that they know to be, in the last analysis, destructive and anti-human. We are building these lethal autonomous robotic weapons but they are going to be "stronger, faster, and smarter than humans." We are going to do better than mere humans and we will save many lives. We believe (or convince ourselves that) we can achieve this chimera and therefore we must try (and, of course, inure ourselves to the burden of accepting the public's money in furtherance of this grotesque illusion.)

Again, I want to insist on the question, "Why should anyone **want** to do this?" In the words of Joseph Weizenbaum, "Technological inevitability can thus be seen to be a mere element of a much larger syndrome. Science promised man power. But, as so often happens when people are seduced by promises of power, the price exacted in advance and all along the path, and the price actually paid, is servitude and impotence. Power is nothing if it is not the power to choose. Instrumental reason can make decisions, **but there is all the difference between deciding and choosing.**"[9, emphasis added] What is it that we are choosing when we choose to develop the ability to make war in a way that is better than the way humans wage war?

6 An Instructive Story

As In mid-October of 1962, photographs taken during a U2 surveillance flight over Cuba revealed the presence of missile sites and Soviet missile components on the island. Assurances given both by Andrei Gromyko, the Soviet Foreign Minister, and Nikita Khrushchev, the leader of the Soviet Union, that no Soviet missiles would be installed in Cuba were thus revealed to be a deception. This precipitated what was in all probability the most dangerous episode of the Cold War, a period of fifteen days in which the two superpowers were on a path to war that would have involved attacks using nuclear weapons by each on the other. The consequences of this were and are unimaginable.

In a chapter of the excellent book, *Humanity: A Moral History of the 20th Century*, Jonathan Glover recounts the story of how Khrushchev and Kennedy managed to step back from the brink in spite of the strong forces – intense military competition, mutual suspicion and misjudgment, internal political pressures, the actions of military subordinates in the forces of both countries that exceeded their standing orders – that tended toward war and nuclear disaster. The conditions surrounding the Cuban Missile Crisis enumerated by Glover recapitulate, in an eerie correspondence, the set of misjudgments, miscalculations, and reckless actions that in 1914 led the European powers into a war that can only be considered a disaster for those who fought and for the generation that survived the conflict. How, then, did the leaders of the two superpowers in 1962 avoid the trap? It is a riveting and illuminating story worth the attention of anyone considering the role of autonomous weapons in war. [10]

The story is riveting because this was a very close call. There were pressures on both leaders – from both the political and military establishments as well as the Cuban leader Fidel Castro – to take actions (including on the U. S. side, an air attack and/or invasion of Cuba) which, with Soviet tactical nuclear weapons already deployed in

Cuba, would almost certainly have led to a catastrophic nuclear exchange. Among the factors that appear to have prevented this, there were two that are worthy of reflection in the context of this paper.

The first is that historian Barbara Tuchman had published earlier that year her study, *The Guns of August*, which carefully dissected European internal political pressures, misunderstandings in regard to treaty commitments, ambiguous signals, poor communication among allies and between potential belligerents, and the military preparations once begun that seemed impossible to roll back that led ineluctably to war and disaster for the continent. Both President Kennedy and his closest advisors (including his brother Robert) had read the book and referred to it during the meetings at which the possible responses to the Soviet threat were discussed. According to the memoirs of Robert Kennedy, quoted in Glover [10], JFK spoke with his brother about the European leaders in 1914 saying "they seemed to tumble into war through 'stupidity, individual idiosyncrasies, misunderstandings, and personal complexes of inferiority and grandeur.' He said, 'I am not going to follow a course which will allow anyone to write a comparable book about this time, The Missiles of October. If anybody is around to write after this, they are going to understand that we made every effort to give our adversary room to move. I am not going to push the Russians an inch beyond what is necessary.'"

Of equal weight, on the Russian side, Khrushchev, early in the crisis, sent a letter to President Kennedy in which he wrote: "Should war indeed break out, it would not be in our power to contain or stop it, for such is the logic of war. I have taken part in two wars, and I know that war ends only when it has rolled through cities and villages, sowing death and destruction everywhere … If people do not display wisdom, they will eventually reach the point where they will clash like blind moles, and then mutual annihilation will commence … You and I should not now pull on the ends of the rope in which you have tied a knot of war, because the harder you and I pull, the tighter this know will become. And a time may come when the knot is tied so tight that the person who tied it is no longer capable of untying it, and then the knot will have to be cut." [10]

Both the words of Nikita Khrushchev and the import of Barbara Tuchman's analysis that was present in the minds of John Fitzgerald Kennedy and his advisors resonated with the warning articulated by Robert E. Lee: "It is good that we find war so horrible, or else we would become fond of it." This was a crisis that both leaders understood would forever indelibly bear their signatures, however it unfolded. That personal sense of responsibility and the consciousness of the horrors of war were the factors that made it possible to pull back. Let us imagine the computer system, designed and implemented by individuals without names and without the wisdom of those who read and reflect and are conscious of the horror, let us indeed pause and imagine the system capable of the saving wisdom of Khrushchev and Kennedy.

7 The Question of Accountability and Responsibility

The "Problem of Many Hands," articulated by Helen Nissenbaum in her 1994 paper [11], has become a common-place, a cliché. We cite this problem by name, nod

knowingly in acquiescence of the certainty that any large software engineering project is bound to have some unanticipated failure modes with serious negative consequences. If, as is customary, the project is developed over a significant period of time by a team the membership of which is not fixed, it will be difficult, perhaps impossible, to determine who is responsible for the failure, to say who should be held accountable for harms ultimately engendered in the use of such a system. This is just a fact of life in our technologically sophisticated world. Get over it and move on.

Perhaps we can agree that there are areas of application where the expectation of future benefits resulting from the development of a new technology justifies accepting the risks of such negative consequences – without, however, relinquishing the understanding that, while we are waiting for the realization of such benefits, someone, some organization must be held accountable and accept responsibility for these harms. There are some areas of application where we can agree to take these risks. But there are assuredly areas where this attitude is unjustifiable. The development of autonomous robotic killing weapons is one of them.

Whose name will be on the disaster precipitated by the predictable malfunction of one of these weapons? Whose name will be attached, as Khrushchev and Kennedy were aware theirs would beto the nuclear disaster precipitated by a reckless gesture in the course of the Cuban Missile Crisis? Who will own the damage to what little of civilized culture we still imagine we possess? Certainly not foolish and opportunistic computer scientists like Arkin, whose names will have long been forgotten. In a sense, this is appropriate. However much their work contributes to this damage, the disaster will be ours as a society if we do not recognize the folly of the path we are taking.

8 Concluding Observations

Finally, it is important to recognize that, although many profound thinkers have contributed to our understanding of what it means to act ethically, our ideas about ethical behavior are as much a product of our experience and our emotional wisdom as of our analytical intelligence. Beware the scientist or engineer who claims that technique will substitute for human instinct and wisdom and enable us to program a machine to behave ethically. Even to approximate this would require the solution of a software engineering problem of forbidding complexity. A moment's reflection on our discouraging experience with such systems should give us pause. [12]

Long ago, Joseph Weizenbaum cautioned against the seduction of technique applied to problems for which its application is utterly inappropriate. "There are two kinds of computer applications that either ought not be undertaken at all, or if they are contemplated, should be approached with utmost caution. …The first kind I would call simply obscene. These are ones whose very contemplation ought to give rise to feelings of disgust in every civilized person. … I would put all projects that propose to substitute a computer system for human understanding for a human function that involves interpersonal respect, understanding, and love in [this] category. [9]

Beyond this, in choosing to invest in the chimerical pursuit of the ability to build machines that can "do better than people" at waging war, we are distorting the

priorities on which a civilized society should rest. We seem unable to make a commitment to educating or providing adequate health care for all the children who live among us, but we find it easy to lavish great sums in the pursuit of an obscenity, oblivious to the warning, "It is good that we find war so horrible, or else we would become fond of it."

References

1. Singer, P. W.,Military Robotics and Ethics: A World of Killer Apps, *Nature*, vol. 477, 22 September,2011, pp. 399-401
2. Singer, P. W., Wired for War: The Future of Military Robots, in *Wired (UK)*, August 2009, available at http://www.brookings.edu/opinions/2009/0828_robots_singer.aspx, last accessed 15 July 2013
3. Singer, P. W., *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin Press, New York, (2009)
4. Singer, P. W., Military Robots and the Laws of War, *The New Atlantic*, Winter 2009 available at http://www.brookings.edu/research/articles/2009/02/winter-robots-singer, last accessed 15 July 2013
5. Singer, P. W., The Ethics of Killer Applications: Why Is It So Hard to Talk About Morality When It Comes to New Military Technology, *Journal of Military Ethics*, vol. 9 no. 4, pp. 299-312 (2010)
6. Iran Air Flight 655, in Wikipedia, at http://en.wikipedia.org/wiki/Iran_Air_Flight_655, , last accessed 15 July 2013
7. Gerdes, A. and Øhrstrom, P., Preliminary Reflections on a Moral Turing Test, in *Proceedings of ETHICOMP 2013, The Possibilities of Ethical ICT*, University of Southern Denmark, Kolding, Denmark, pp. 167-174 (2013)
8. Parnas, D. L, Letter to James H. Offutt, in *Introduction to Computer Ethics, Parts 1 and 2*, at *www.stanford.edu/class/cs181/materials/CS181-Parts1and2.pdf*, last accessed 15 July 2013
9. Weizenbaum, Joseph, *Computer Power and Human Reason: From Judgment to Calculation*, W. H. Freeman, New York, (1976)
10. Glover, Jonathan, *Humanity: A Moral History of the 20th Century*, *2nd edition*, Yale University Press, New Haven, (2012)
11. Nissenbaum, H., Computing and Accountability, Communications of the ACM, vol. 37, no. 1, pp. 73-80 (1994)
12. Fleischman, W., Electronic Voting Systems and the Therac-25: What Have We Learned?, in , *Proceedings of ETHICOMP 2010, The "Backwards, Forwards, and Sideways Changes" of ICT*, UniversitatRoviraiVirgili, Tarragona, Spain, pp. 170-179 (2010)

Declarado de Interés Municipal por el Honorable Concejo
Deliberante del Partido de General Pueyrredon

Secretaría de Políticas Universitarias

CONICET

UNIVERSIDAD CAECE Mar del Plata

RedUNCI