

RESEARCH IN COMPUTING SCIENCE

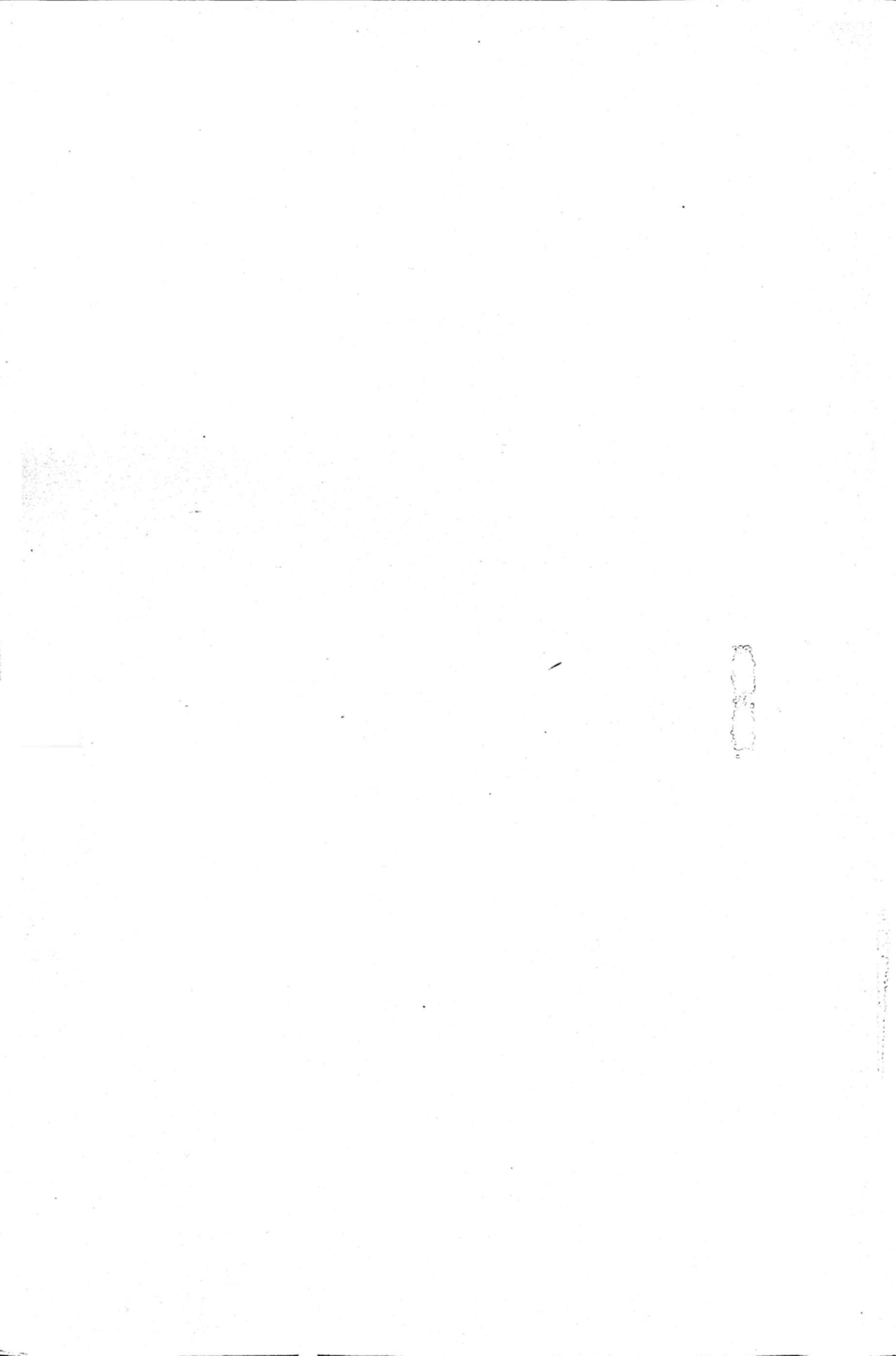
ISSN: 1870-4069

Advances in Computer Science and Engineering

Miguel Martínez
Antonio Alarcón
(Eds.)

Vol. 45

RCS
Research in Computing Science



Advances in Computer Science and Engineering

Research in Computing Science

Series Editorial Board

Comité Editorial de la Serie

Editors-in-Chief:

Editores en Jefe

Juan Humberto Sossa Azuela (Mexico)
Gerhard Ritter (USA)
Jean Serra (France)
Ulises Cortés (Spain)

Associate Editors:

Editores Asociados

Jesús Angulo (France)
Jihad El-Sana (Israel)
Jesús Figueroa (Mexico)
Alexander Gelbukh (Russia)
Joannis Kakadiaris (USA)
Serguei Levachkine (Russia)
Petros Maragos (Greece)
Julian Padget (UK)
Mateo Valero (Spain)

Editorial Coordination:

Coordinación Editorial

Blanca Miranda Valencia

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 45**, Mayo 2010. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. 04-2004-062613250000-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de Licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor Responsable: *Juan Humberto Sossa Azuela, RFC SOAJ560723*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 45**, May, 2010. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, May, 2010, in the IPN Graphic Workshop – Publication Office.

Volume 45

Volumen 45

Advances in Computer Science and Engineering

Volume Editors:

Editores del Volumen

Miguel Martínez

Antonio Alarcón

Instituto Politécnico Nacional
Centro de Investigación en Computación
México 2010



ISSN: 1870-4069

Copyright © Instituto Politécnico Nacional 2010

Copyright © by Instituto Politécnico Nacional

Instituto Politécnico Nacional (IPN)

Centro de Investigación en Computación (CIC)

Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal

Unidad Profesional "Adolfo López Mateos", Zacatenco

07738, México D.F., México

<http://www.ipn.mx>

<http://www.cic.ipn.mx>

The editors and the Publishers of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

All rights reserved. No part of this publication may be reproduced, stored on a retrieval system or transmitted, in any form or by means, including electronic, mechanical, photocopying, recording, or otherwise, without prior permission of the Instituto Politécnico Nacional, except for personal or classroom use provided that copies bear the full citation notice provided on the first page of each paper.

Indexed in LATINDEX

Indexada en LATINDEX

Printing: 500

Printed: 500

Printed in Mexico

Impreso en México

Preface

The purpose of this volumen is to reflect the new directions of investigation in selected áreas of Computer Science, mainly related with artificial intelligence. Papers for this volumen were carefully selected by volumen editors on the basis of the blind reviewing process performed by editorial board members and additional reviewers. The main criteria for selection were their originality and technical quality.

This issue of the journal Research in Computing Science can be interesting for researchers in computer science, especially in áreas related to artificial intelligence, and also for persons who are interested in cutting edge themes of the computer science.

This volumen contains revised version of 28 accepted papers by 78 authors, selected for publication after thorough evaluation process.

The papers are structured into the following sections:

- Algorithm Theory.
- Neuronal Networks and Optimization.
- Image Processing.
- Neuronal Language Processing and Knowledge based Systems.
- Robotics.
- Computer Networks and Web Services.
- Software Engineering.
- Artificial Intelligence.

This value is a result of work of many people. In the first place, we thank the authors of the papers included in this volumen for the technical excellence of their papers that assures the high quality of this publication.

We also thank the members of the International Editorial Board of the volumen and the additional reviewers for their hard work consisting in selection of the best papers out of many submissions that were received.

We would like to thank to personnel of the Center for Research in Computer Science of the National Polytechnic Institute, for their indispensable help in the process of preparation of the volumen.

May, 2010

Miguel Martínez
Antonio Alarcón

The
proceedings of
the

Table of Contents

Índice

	Page/Pág.
Analysis and Risk Assessment of the Real Time Volcanic Monitoring System.....	3
Luis Enrique Colmenares Guillén and Omar Ariosto Niño Prieto	
Building a Minimal Spanning Tree for the #2SAT Problem.....	15
Guillermo De Ita, Meliza Contreras, Pedro Bello	
A Hybrid Evolutionary Algorithm for the Edge	
Crossing Minimization Problem in Graph Drawing.....	27
Sergio Enríquez, Eunice Ponce de León, Elva Díaz and Alejandro Padilla	
A New Approach to Music Information Retrieval using Dynamic Neuronal Networks.....	41
L.E. Gomez, J.H. Sossa, R. Barron, J.F. Jimenez	
Conformal geometric algebra and sphere fitting applied to colour image segmentation.....	53
Luis Horna, Ricardo Barrón and Giovanni Guzmán	
Algorithm of support for the detection of the Acute Lymphoblastic Leukemia.....	59
Susana Ordaz Gutiérrez, Fabián Torres Robles,	
Francisco Javier Gallegos Funes, Alberto Jorge Rosales Silva.	
Formal Verification for the Absence of Deadlock in the Manager Workers Pattern.....	73
Jorge Luis Ortega-Arjona and Francisco Hernández-Quiroz	
Medical Carnet for Management of Patients Driven by Ontologies.....	85
F. Mata and S. Zepeda	
Design of a High Dynamic Range ADC by Concatenating Low Resolution Samples.....	97
Miguel Santiago Villafuerte Ramírez, Alfonso Gutiérrez Aldana	
and Luis Pastor Sánchez Fernández	
A Software Tool for the Analysis of Similarity in Recurrence Patterns.....	109
Ernesto Bautista-Thompson, Roberto Brito-Guevara, Jesús E. Molinar-Solis	
Intelligent and Adaptive User Interfaces for Ubiquitous Learning.....	119
Héctor Antonio Villa Martínez, Francisco Javier Tapia Moreno	
Using the Software Process Improvement approach for defining a Methodology for Embedded Systems Development using the CMMI-DEV v1.2.....	127
García, I. and Herrera, A.	
Population Coding and SpikeProp Hardware Accelerator for Spiking Neural Networks.....	145
Marco Aurelio Nuño-Maganda, Cesar Torres-Huitzil, and Miguel Arias-Estrada	
An Intelligent Virtual Agent for Collaborative Learning looking to be part of the Team.....	157
Raúl A. Aguilar, Angélica de Antonio, Ricardo Imbert and Adriana Peña	
Distributed System for Assessment of Water Quality in Shrimp Aquaculture Systems.....	169
José Juan Carbajal Hernández, Raúl A. Valero Cruz	
and Mauricio Suárez López	

Intelligent Fault Diagnosis and Prognosis using state validations in a drinking water plant.....	179
Hector Hernandez, Jorge Camas, Nicolás Juárez, Madaín Pérez, Rafael Mota1 and Claudia Isaza	
Web-Mapping Application to Retrieve Spatial Data by means of Spatial Ontologies.....	191
Miguel Torres, Marco Moreno, Rolando Quintero and Giovanni Guzmán	
Applying Diverse Data Mining Methods in the Electric Power Industry.....	209
Manuel Mejía-Lavalle, Guillermo Rodríguez O., Gustavo Arroyo F. and Eduardo F. Morales	
Using WEKA for Semantic Classification of Spanish Verb-Noun Collocations.....	221
Olga Kolesnikova and Alexander Gelbukh	
Experimenting with Maximal Frequent Sequences for Multi-Document Summarization.....	233
Yulia Ledeneva, René Arnulfo García-Hernández, Anabel Vazquez-Ferreira, Nayely Osorio de Jesús	
Establishing a Software-Subcontracting Management Model to Improve the Software-Subcontracting Process in Small-size Enterprises.....	245
García, I. and Pacheco, C.	
Supporting the Management Process of Software Process Improvement Initiatives based on NMX-I-059/02-NYCE-2005.....	261
García, I. and Cruz, D.	
Security and Adaptability to Groupware Applications using a Set of SOA-based Services.....	279
Mario Anzures-García, Luz A. Sánchez-Gálvez, Miguel J. Hornos, and Patricia Paderewski-Rodríguez	
Using the CPAN Branch & Bound for the Solution of Travelling Salesman Problem.....	291
Mario Rossainz López, Manuel I. Capel Tuñón	
Fast Automatic Retinal Blood Vessel Segmentation and Vascular LandmarksExtraction Method for Biometric Applications.....	303
Fabiola M. Villalobos-Castaldi, Edgardo M. Felipe-Riverón	
Implementation of a swarm intelligence algorithm to a mobile device.....	317
L.E. Gomez, J.F. Jimenez, J.H. Sossa, F.J. Cuevas, O. Pogrebnyak, R. Barrón	
Demodulation of a single Interferogram by use a Parametric Method based on a Differential Evolution.....	327
J.F. Jimenez, F.J. Cuevas, J.H. Sossa, L.E. Gomez	
Enhancing the Diagnosis Module in a Self-healing Architecture Supporting Web Service Applications.....	337
Francisco Moo-Mena, Fernando Curi-Quintal, Juan Garcilazo-Ortiz, Luis Basto-Díaz, and Roberto Koh-Dzul	
Author Index.....	349
Índice de autores	
Editorial Board of the Volume.....	351
Comité editorial del volumen	

Analysis and Risk Assessment of the Real Time Volcanic Monitoring System

Luis Enrique Colmenares Guillén¹, Omar Ariosto Niño Prieto^{1,2}

¹
Benemerita Universidad Autonoma de Puebla,
Facultad de Ciencias de la Computación,
BUAP – FCC, Ciudad Universitaria,
Apartado Postal J-32,
Puebla, Pue. México.
lecolme, omar.ariosto@gmail.com

²
Université Claude Bernard Lyon I
Bâtiment Nautibus
43, Boulevard du 11 novembre 1918
69622 Villeurbanne Cedex France
OMAR.NINO-PRIETO@bvra.etu.univ-lyon1.fr, omar.ariosto@yahoo.fr

Abstract. In this work, first present the methodology used in the Real Time Volcanic Monitoring System such as SA-RT (Structured Analysis for Real Time) used for the design of the system, and the system itself created for the prevention of the consequences during a catastrophic volcanic event. During the second phase, the Analysis and Risk Assessment of the system is presented. The main contribution of this paper is the complete Risk Assessment Analysis done after finished the design of the system. In order to make a complete Risk Assessment Analysis for complex systems and critical systems done by the engineers, some methodologies are used like the Fault Tree Analysis (FTA), The Markov Analysis and the Petri Nets. In this paper all of them are presented and used since the beginning. First, the Petri Nets are used by the SA-RT methodology in the design of the system. Then the Fault Tree Analysis is used to perform the reliability and safety analysis and to prevent the consequences of an eventually catastrophic volcanic event. The Markov Analysis is used to model all the system because it has high dependencies between its components. Finally the parallel software design is done by the LACATRE methodology.

Keywords: Risk Assessment, Qualitative Risk Assessment, Quantitative Risk Assessment, Markov Analysis, Fault Tree Analysis.

1 Introduction

The active volcanoes are a risk for the society that live near, so it is necessary to study them in order to predict a major event and prevent its consequences like the lost of lots of human lives, the environmental damage and in general the society is very

affected because of this. The Software and Hardware systems help the specialists to predict and to understand this activity better than before, but still doesn't exist 100% of true prediction about a catastrophic volcano event. The Risk Assessment Engineering of certain complex systems[13, 14] like the critical systems of a nuclear plant helps to prevent catastrophic events[11, 14, 15] but in the case of natural disasters exist event a more random probability to occur because is not in function of human activity.

The volcanic monitoring is a hard work that requires a lot of time and effort to have success. So it is necessary to use an automatic system to do this job but also it is necessary to make the Risk Assessment Analysis in order to maintain it always working. The Real Time Monitoring System is a developer tool that helps to monitoring the Popocatepetl activity, additionally of the existing monitoring systems like CENAPRED systems UNAM system and the related work done by many countries in order to prevent this catastrophic event [1, 2].

2 Real Time Monitoring Systems

2.1 Architecture Requirements

Two magnetometers are located strategically in the cone of the Popocatepetl volcano to monitor the magnetic activity and they would transmit that activity through a GSM Web that will transmit the data by Internet [1]. A computer system located in the research center of CUPREDER will receive that information and the specialists will be able to make an interpretation of the actual activity, but some Critical Parameters shall be taken into consideration in order to register an abnormal activity and try to predict the comportment of the volcano[1,2] (Fig. 1).

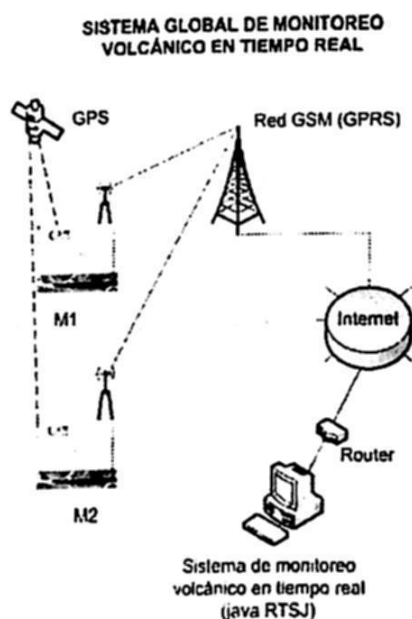


Fig. 1. Global System

2.2 Real Time Monitoring System Methodology

The methodology used to design the Real Time Monitoring System was SA-RT (Structured Analysis for Real Time) that shows a coherent and structured vision of the design of Real Time Systems. The objective of the magnetometers is to transmit the magnetic signal by wireless information to the GSM system that will send that wireless signal to the RTSJ System. While the system is working, the information will be shown in the screen and if there should be an abnormal activity registered by the critical parameters an alarm will be activated. All the data will be saved in a Database that will be accessed by another analysis subsystem [1, 2, 3, 4, 6 and 8]. This is shown by the context diagram (Fig. 2).

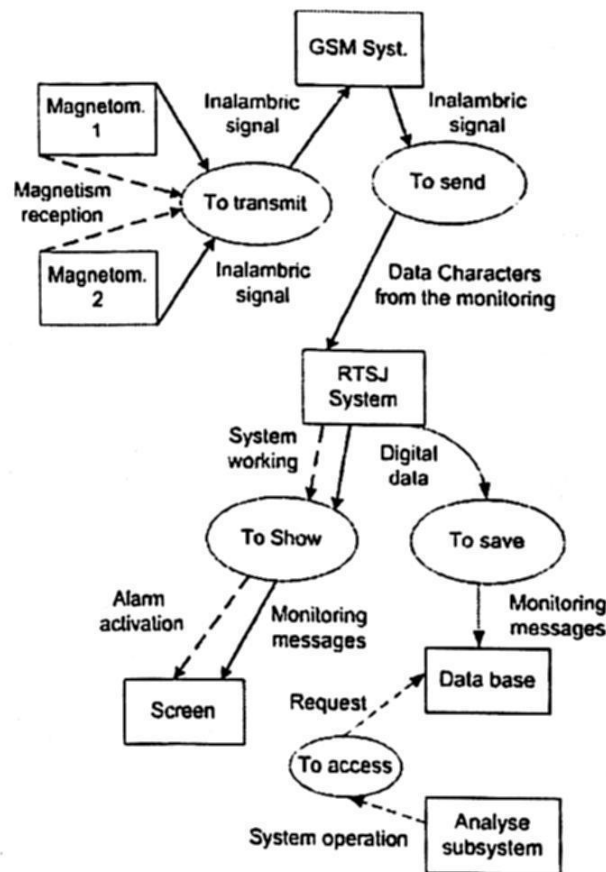


Fig. 2. Context diagram

The Data Flow Diagram (DFD) is shown below (Fig. 3) with all the system processes and the events that enter and go out from the control bar. Also the data flow interact with the whole system and can be compared with the Critical Parameters in order to compare and better understand an abnormal volcanic activity if produced. The State Transition Diagram is derived from the DFD (Fig. 4) and shows each state of the system while functioning.

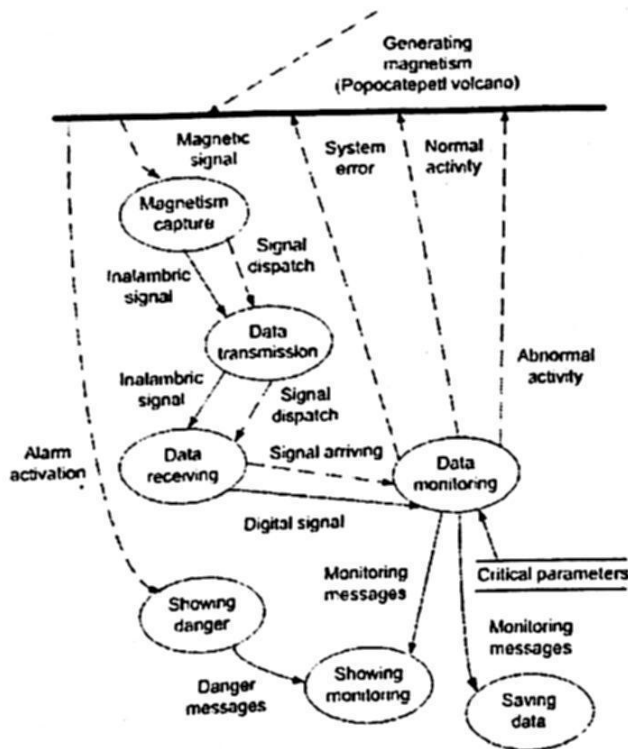


Fig. 3. Data Flow Diagram (DFD)

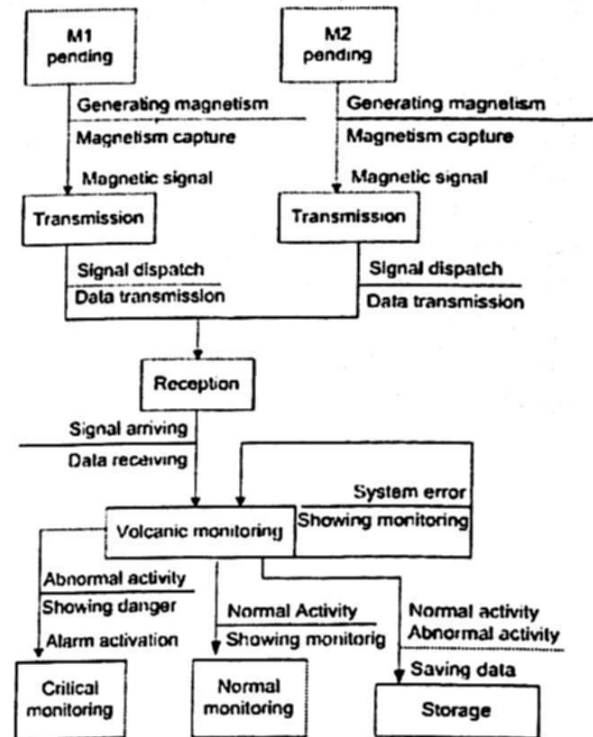


Fig. 4. State Transition Diagram

3 An approach of Analysis and Risk Assessment of the system

The Risk Assessment is a discipline that search the probability of catastrophic events that have catastrophic consequences such as the lost of human lives or irreversible economical and environmental damage [7]. For maintaining the components of a system in a good state in order to prevent those consequences, the Analysis of Risk Assessment of a system is made by the Qualitative Analysis and the Quantitative analysis [10]. The *Qualitative risk assessment* requires calculations of two components of risk: R , the magnitude of the potential loss L , and the probability p , that the loss will occur. The *Quantitative risk assessment* determines the probability of the occurrence of a catastrophic event, and the weakness of a system [10, 11].

The availability is the probability of the good operation of a component. The reliability is the capacity of success of a component during a period of time. The reliability study is made to ensure the success of critical systems that works in cold or warm redundancy [10, 11, and 20].

3.1 The mathematical models

There are some mathematical models used in the Analysis and Risk Assessment of the systems. One of the most used is the Fault Trees, The Markov Analysis, and the Petri Nets [10, 11, 12 and 20] which are used in the design of the whole system.



Fig. 5. Mathematical models

3.2 Analysis with a Fault Tree

This method construct a logic connected diagram by AND and OR gates It has the objective to find the combinations of failures of the components and the *minimal cut set* that describe the combinations of component failures that cause the TOP catastrophic event to occur. This method is deductive and it has the top event, intermediate event and the base event which is the beginning of the failure of the system [10, 11, 12 and 20]. For the Analysis and Risk Assessment of the Real Time Volcanic System, the following Fault tree is proposed:

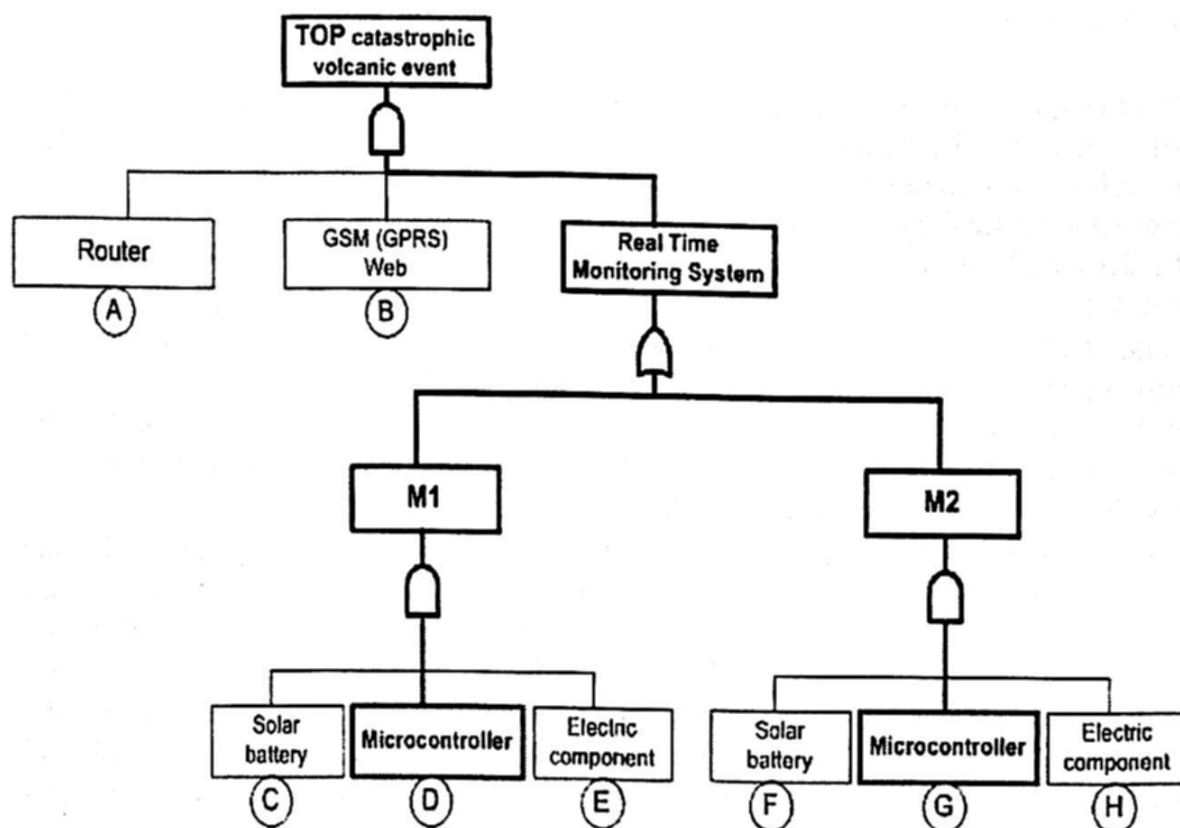


Fig. 6. Real Time Monitoring System Fault Tree

In order to understand the fault tree, first the base failures like the failure of the micro-controller or the battery or any electric component of the design of the system [1] are introduced. After those failures, one of the magnetometers would send corrupted data if the micro-controller fails or simply the magnetometers (M1, M2) stop working. The design of the system allows working in warm redundancy. If one of the magnetometers fails, the system could still be working but with less data to provide to the specialists during the volcanic event. The system fails if the Real Time System fails, because the two magnetometers don't have any sense without an interpretation of the data received by them. Nevertheless, if the Real Time System fails, the data from the magnetometers could be taken manually during normal conditions where there is not an important volcanic event. This doesn't work if exist a catastrophic volcanic event. The minimal cut set is bold in the Real Time Monitoring System Fault Tree, and the mathematical Boolean expression of this tree in order to express the TOP event is represented like follows [10, 11, 12, and 20]:

$$TOP = A * B * ((C * D * E) + (F * G * H))$$

Each base event (A, B, C, D, E, F, G, and H) has a probability $p(x)$ to be produced.

3.3 Markov Analysis

The Markov Analysis permits to visualize the states of the systems and the transitions between them. This method visualizes diagrams of state and space of the systems behavior. This method makes a detailed analysis of the systems that could have another intermediary state different from failure or success. It also permits to study the degraded systems and a probability is associated for each state changing [10, 11, and 20].

The Markov modeling is not random because the states are dependent from the last immediate state but independent from the others [10]. This can model systems without memory and the probability of changing of state is constant. The comportment of the system depends on the present state that is continuous in the time and discrete in space [10, 11, and 20].

The Real Time Volcanic Monitoring System has different components like the magnetometers (A, B) which has the comportment of repairable systems, the Real Time RTSJ (C) that has the comportment of degradable system and they are represented by the Chains of Markov in Alta Rica Data Flow Programming [16, 17, 18, 19, 21 and 22]. Those models were created before in order to understand, simulate and predict the comportment of the critical systems [17, 18, 19, 21 and 22], and adapted for the Real Time Volcanic Monitoring System components. Those models could be more complex, but the basic representation for the correct operation is represented below.

Component A:
Magnetometer 1 (M1)

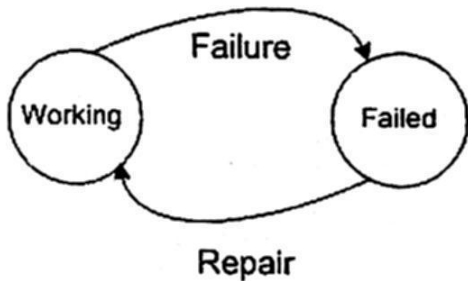


Fig. 7. Repairable model A

Component B:
Magnetometer 2 (M2)

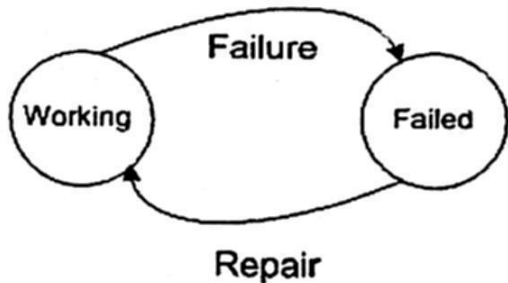


Fig. 8. Repairable model B

Component C:
Real Time System Model

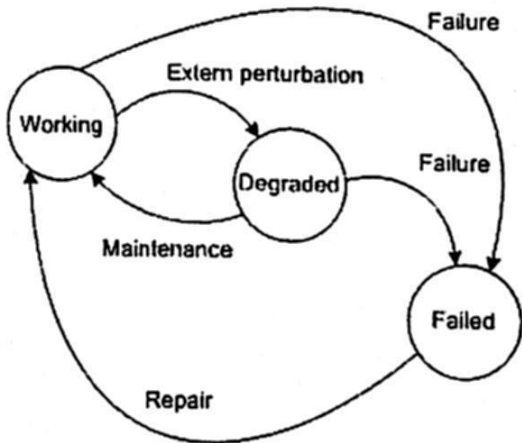


Fig. 9. Degraded System model C

Each component has a compartment Markov model, but the interaction of the system is done by the composition [16, 17, 18, 19, 21 and 22] of the models of the basic components. First twelve different combinations of the states of the components could be found, but in order to find the correct model of the system, the semantic of The Real Time Volcanic Monitoring System is respected. The concrete model of the Real Time Volcanic Monitoring System is proposed below.

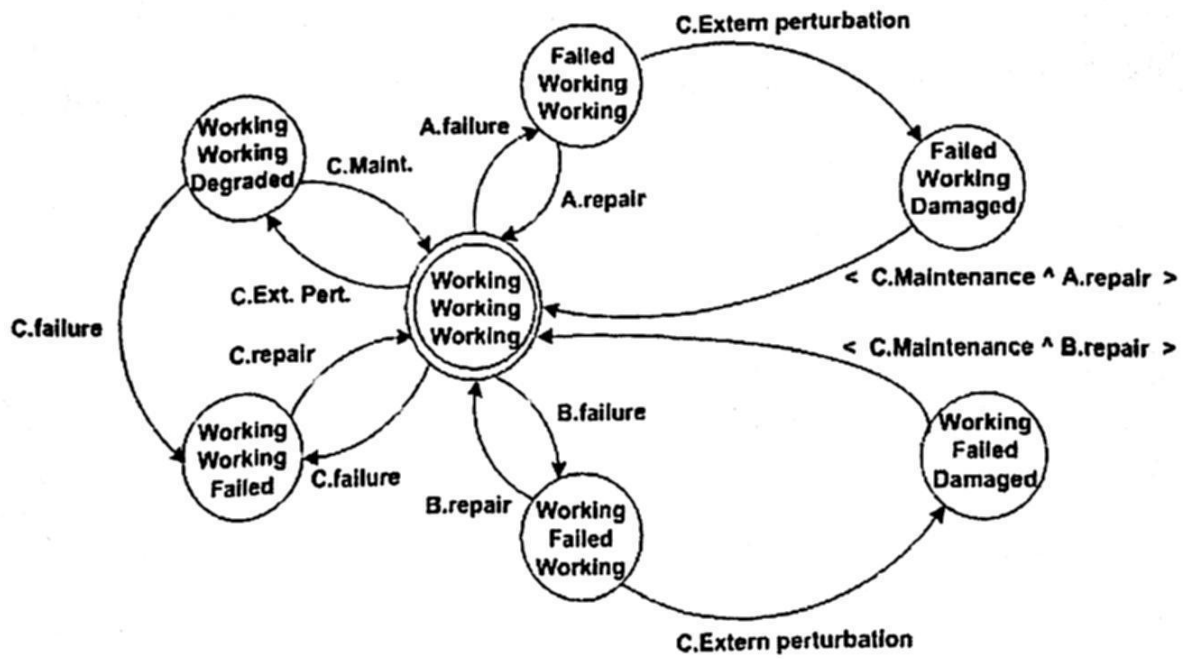


Fig. 10. The Real Time Volcanic Monitoring System Model

The correct operation of the Real Time Volcanic Monitoring System doesn't allow that the three components fail at the same time, because if exist a volcanic event, the system wouldn't be able to report a volcanic abnormal event nor predict it [2]. So there are some states that don't have any sense to stay. Also the system should be in warm redundancy [15] because there are two Magnetometers working at the same time and providing important data while they are transmitting the magnetism [1]. In the beginning, the whole system is working; if exist a failure from the magnetometer one or two the state will be changed, but the system will continue working. After this state, the computer system could be damaged or have any external perturbation so there is a new changing of state. At this time, the system is still working but it has to return to the initial state to continue working without any problem, so there are transitions that are mixed up with the synchronization [16, 17, 18, 19, 21 and 22] in order to return as soon as possible to the initial state. The computer system could be damaged and eventually fail while the two magnetometers are working this has a new state and this is also modeled. Finally, the prevention is the main idea of the model of the system, knowing that the TOP event consequence (human, economical) is the lack of prevention and incertitude of a major volcanic event.

4 The software design of the system

The software design [1] is represented by the LACATRE [5, 9] real time systems methodology (Fig. 11). The main program is divided in several modules. The main () function will work with the following threads:

- P2: Magnetometer 1, with priority 1.
- P3: Magnetometer 2, with priority 2.
- P4: Monitoring, with priority 3.
- P5: Alarm, with priority 4.
- P6: Prediction, with priority 5.

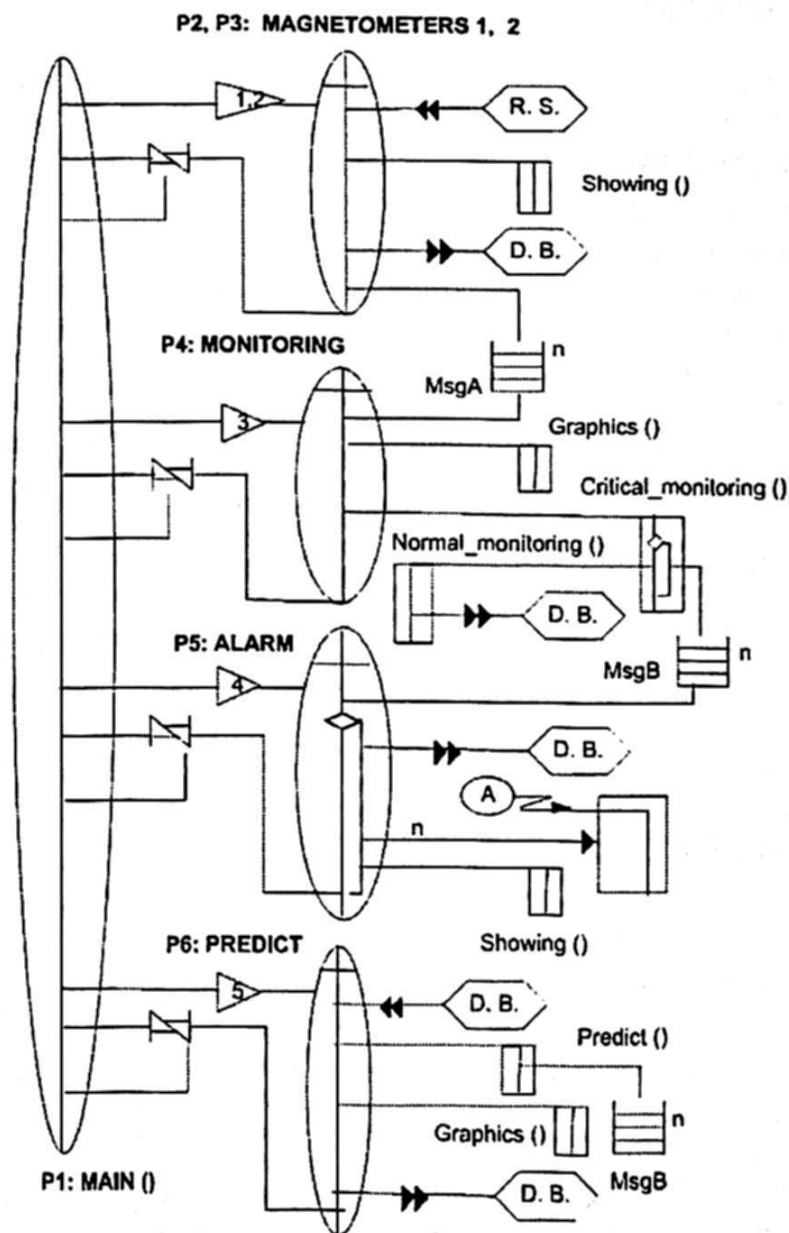


Fig. 11. LACATRE Diagram

The threads are synchronized by semaphores to ensure that all the processes have the correct data. The signal from the magnetometers are obtained with a communication data protocol and transmitted by the data resource (R.S.) to the threads P1 and P2 that shows the magnetic activity with the function *Showing()* that has all the procedures to show the numerical data in the screen. Then all the data are saved in a Data Base and the messages from each magnetometer are sent independently to one FIFO data structure, and the thread P4 receive them to process the information received by the functions *Graphics()*, *Critical_Monitoring()* and *Normal_Monitoring()*.

If an abnormal activity exists, the function *Critical_Monitoring()* send a message to another FIFO to register the abnormality and those messages are transmitted to the thread P5 which write the abnormal data in the data base and activate an alarm. At the same time the function *Showing()* is showing all the abnormal activity in the screen. The system has another thread P6 which has the function *Prediction()*, and this function makes a prediction of the magnetic activity using neural networks algorithms [2].

5 Conclusions and Ongoing Research

The main contribution of this work is to have an approach of the risk assessment of the Real Time Volcanic Monitoring System. The catastrophic event is not dependent from human activity like other types of critical systems but always exist a probability of occurrence. Nevertheless if we use the same methodology and the catastrophic event is produced, we will be able to reduce the consequences and the people will be more prepared to face and solve the problems derived from this event like human lives and economical damage. The human factor is always important because sometimes the natural catastrophic events take some time to be produced but it is very important to be prepared.

At this time for the project, many subjects are being studied; one of them the possibility to implement all the system but it is necessary to work as a team with many specialists in electronics, computing science, geologists and volcanologist, all of them from Puebla working as a team at the University of Puebla BUAP, and CUPREDER [1, 2]. This project will help to prevent the natural disasters very frequent in Mexico. We are not able to avoid them, but reduce the negative consequences if happened.

References

1. O. Niño, E. Colmenares, M. Martin, Sistema de Monitoreo Volcánico en Tiempo Real Congreso de IEEE, ANDESCON2008 (Ciudad del Cusco, Peru) Octubre de 2008. ISBN: 978-603-45345-0-6.
2. O. Niño, E. Colmenares, M. Martin, Propuesta para predecir eventos en el Sistema de Monitoreo Volcánico en Tiempo Real basado en la informática bio-inspirada y algoritmos de Redes Neuronales. Congreso SENIE Aguascalientes

- 2008 Universidad Autónoma de México. (UAM) Oct. de 2008. ISBN : 978-970-31-0944-9
3. BABAU Jean-Philippe, cours 4IF "Conception et Integration d'Applications Industrielles, SA-RT Structured Analysis for Real-Time 4IFCIAI" (Annee 2007-2008), INSA de Lyon, Couse notes.
4. BABAU Jean-Philippe, cours 4IF "Systemes d'Exploitation avances 4IFSEA, conception Multitâches Lacatre/VxWorks", (2007-2008), INSA de Lyon, Couse notes.
5. Piotr Szwed, Lacatre Reference Guide, programming tool, http://pszwed.ia.agh.edu.pl/RT/La4_rm/La4_rm.html (2008).
6. Prih Hastono and Sorin A. Huss, Automatic Generation of ExecutableModels from Structured Approach Real-Time Specifications <http://www.vlsi.informatik.tu-darmstadt.de/staff/hastono/rtss04-sart.pdf> 2008.
7. J.A. McCall, Factors in Software Quality, General Electric no. 77C1502, jun 1977.
8. Babau, Jean-Philippe, SA-RT "Structured Analysis for Real-Time" <http://www.if.insa-lyon.fr/chercheurs/jpbabau/cours/sart.pdf> (2008).
9. J. J. SCHWARZ, J. J. SKUBICH, Graphical programming for Real-Time Systems, Control Engineering. Practice, Vol. 1, No. 1, pp. 43-49, 1993.
10. Marvin Rausand ; System Reliability Theory, Models and Statistical Methods; John Wiley & Sons, Inc. 1994.
11. J.D Andrews and T.R. Moss; Reliability and Risk Assessment ; Longman Scientific & Technical; 1993.
12. <http://faulttreesoftware.info/>(2009)
13. http://www.ixxi.fr/Les_SC.php (2009)
14. Ingénierie système, notes de cours M2, ingénierie des systèmes complexes; Thales Université-Ecole Polytechnique, France ; 2007-2008.
15. Storey Neil, Safety-Critical Computer Systems; Addison Wesley; Longman; 1996.
16. Rauzy Antoine, Griffault Alain, Point Gérald, Arnold André ; The Alta Rica Formalism for Describing Concurrent Systems ; Université Bordeaux I, CNRS; 2000.
17. Pont Gérald ; Thèse: Alta Rica Contribution à l'unification des méthodes formelles et de la sûreté de fonctionnement ; Université Bordeaux I ; 2000.
18. Roux Olivier, Pagetti Claire, Cassez Frank ; A Timed Extension for Alta Rica; CNRS Nantes ; 2002.
19. Rauzy Antoine, Alta Rica a Formal Language for Oriented Modeling ; IML/CNRS & ARBoost Thechnologies, Marseille, France ; notes de cours.
20. Gondran M. Pagès A. Fiabilité des Systèmes ; Direction des Etudes et Recherches d'Electricité de France ; 1980.
21. <http://www.pdfgeni.com/book/AltaRica-pdf.html> (2010).
22. Pagetti Claire, Une extension temporisée d'AltaRica Application à la modélisation d'un système embarqué ; IRCCyN 1 rue de la Noë BP 92101, 44321 Nantes Cedex 3 France.

Building a Minimal Spanning Tree for the #2SAT Problem

Guillermo De Ita, Meliza Contreras, Pedro Bello

Faculty of Computer Science, Universidad Autónoma de Puebla
{deita,mcontreras,pbello}@cs.buap.mx

Abstract. Due to #2SAT is a #P-complete problem, different efficient alternatives have been proposed for approximate solutions to #2SAT. We exploit the existent relation between counting models for two conjunctive forms (2-CF's) and Fibonacci numbers that allow us to count the number of models of the Boolean formula in an incremental way.

We design a polynomial time algorithm for given a 2-CF Σ , to build its constrained graph G_Σ and a spanning tree A_Σ such that $\#SAT(A_\Sigma)$ has a minimal number of models into the set of all spanning tree of G_Σ .

Keywords: Counting the Number of Models, Enumerative Combinatorics, Minimal Spanning Tree.

1. Introduction

Counting combinational objects over graphs has been an interesting and important area of research in Mathematics, Physics, and Computer Sciences. The counting problems, being mathematically interesting by themselves, are closely related to important practical problems. For instance, reliability issues are often equivalent to counting problems. Computing the probability that a graph remains connected given the probabilities of failure over each edge is essentially equivalent to counting the number of ways in which those edges could fail without losing connectivity [2], [8].

Due to #2SAT is a #P-complete problem [5], [3] different efficient methods have been developed for counting, although approximately, the number of models for Boolean formulas in two Conjunctive Forms (2-CF) and since #2SAT is a key problem to clarify the frontier between efficient counting and intractable counting procedures [4].

Let Σ be a 2-CF and G_Σ its connected constrained graph. The combinatory problem that we address here is to approximate the number of models for Σ through to build in polynomial time, a spanning tree A_Σ from G_Σ and at the same time to compute the value $\#SAT(A_\Sigma)$ holding:

1. $\#SAT(A_\Sigma) \geq \#SAT(\Sigma)$
2. $\#SAT(A_\Sigma)$ is minimal into the set of all spanning tree of G_Σ

There are some observations about the values: $\#SAT(\Sigma)$ and $\#SAT(A_\Sigma)$. To identify if $\#SAT(\Sigma)$ is zero or greater or equal to one can be done in polynomial time since the 2SAT problem is solved in polynomial time. If $\#SAT(\Sigma) > 1$ then as far as we know, there is not polynomial time algorithm for computing $\#SAT(\Sigma)$. All spanning tree of G_Σ represent formulas where their number of models is greater than 1. The unique way that a 2-CF Σ represents an unsatisfiable formula is that G_Σ contains cycles.

The techniques for building minimal spanning trees have been developed assuming static weights on the edges of the graph [7]. But when we are considering the #2SAT problem, instead of static weights we have dynamic weights determined by the signs of each edge, as well as the number of partial models associated with the endpoints of the edge.

We address the construction of a minimal spanning tree of a constrained signed graph based on the partial values computed in each node of the constrained graph and in the signs of its edges, determining so a new way to build spanning trees with dynamic weights in the edges of the graph.

2. Preliminaries

Let $X = \{x_1, \dots, x_n\}$ be a set of n boolean variables. A *literal* is either a variable x_i or a negated variable \bar{x}_i . As usual, for each $x_i \in X$, $x_i^0 = \bar{x}_i$ and $x_i^1 = x_i$.

A *clause* is a disjunction of different literals (sometimes, we also consider a clause as a set of literals). For $k \in \mathbb{N}$, a *k-clause* is a clause consisting of exactly k literals and, a $(\leq k)$ -*clause* is a clause with at most k literals. A variable $x \in X$ appears in a clause c if either x or \bar{x} is an element of c .

A *conjunctive form* (CF) F is a conjunction of clauses (we also consider a CF as a set of clauses). We say that F is a *monotone CF* if all of its variables appear in unnegated form. A *k-CF* is a CF containing only k -clauses and, $(\leq k)$ -CF denotes a CF containing clauses with at most k literals. A *k μ -CF* is a formula in which no variable occurs more than k times. A $(k, j\mu)$ -CF ($(\leq k, j\mu)$ -CF) is a k -CF ($(\leq k)$ -CF) such that each variable appears no more than j times.

We use $v(X)$ to express the variables involved in the object X , where X could be a literal, a clause or a CF. For instance, for the clause $c = \{\bar{x}_1, x_2\}$, $v(c) = \{x_1, x_2\}$. $Lit(F)$ is the set of literals appearing in F , i.e. if $X = v(F)$, then $Lit(F) = X \cup \bar{X} = \{x_1, \bar{x}_1, \dots, x_n, \bar{x}_n\}$. We denote $\{1, 2, \dots, n\}$ by $\llbracket n \rrbracket$.

An assignment s for F is a boolean function $s : v(F) \rightarrow \{0, 1\}$. An *assignment* can also be considered as a set of non complementary pairs of literals. If $l \in s$, being s an assignment, then s turns l *true* and \bar{l} *false*. Considering a clause c and assignment s as a set of literals, c is *satisfied* by s if and only if $c \cap s \neq \emptyset$, and if for all $l \in c$, $\bar{l} \in s$ then s falsifies c .

If $F_1 \subset F$ is a formula consisting of some clauses from F , and $v(F_1) \subset v(F)$, an assignment over $v(F_1)$ is a *partial* assignment over $v(F)$. Assuming $n = |v(F)|$ and $n_1 = |v(F_1)|$, any assignment over $v(F_1)$ has 2^{n-n_1} extensions as assignments over $v(F)$.

Let F be a CF, F is *satisfied* by an assignment s if each clause in F is satisfied by s . F is *contradicted* by s if any clause in F is contradicted by s . A model of F is an assignment for $v(F)$ that satisfies F . The SAT problem consists of determining if F has a model and $\text{SAT}(F)$ denotes the set of models of F . The #SAT problem (or #SAT(F)) consists of counting the number of models of F defined over $v(F)$. #2-SAT denotes #SAT for formulas in 2-CF. We also denote #SAT(F) by $\mu_{v(F)}(F)$ or just $\mu(F)$ when $v(F)$ is clear from the context.

Let Σ be a 2-CF, the *constrained graph* of Σ is the undirected graph $G_\Sigma = (V(\Sigma), E(\Sigma))$, with $V(\Sigma) = v(\Sigma)$ and $E(\Sigma) = \{\{v(x), v(y)\} : \{x, y\} \in \Sigma\}$, i.e. the vertices of G_Σ are the variables of Σ , and for each clause $\{x, y\}$ in Σ there is an edge $\{v(x), v(y)\} \in E(\Sigma)$.

Each edge has associated an ordered pair (s_1, s_2) of signs, assigned as labels. For example, the signs s_1 and s_2 for the clause $\{\bar{x} \vee y\}$ are related to the signs of the literals x and y respectively, then $s_1 = -$ and $s_2 = +$ and the edge is denoted as: $x \overset{-}{\pm} y$ which is equivalent to $y \overset{+}{\pm} x$.

A graph with labeled edges on a set S is a pair (G, ψ) , where $G = (V, E)$ is a graph, and ψ is a function with domain E and range S . $\psi(e)$ is the label of the edge $e \in E$. Let $S = \{+, -\}$ be a set of signs. Let $G = (V, E, \psi)$ be a signed graph with labelled edges on $S \times S$. Let x and y be nodes in V . If $e = \{x, y\}$ is an edge and $\psi(e) = (s, s')$, then s (s') is called the *adjacent sign* of x (y).

Let $G_\Sigma = (V, E)$ be a constrained graph of a 2-CF Σ . Sometimes, $V(G)$ and $E(G)$ are used to emphasize the graph G . We denote the cardinality of a set A by $|A|$.

The neighborhood of a vertex $v \in V$ is the set $N(v) = \{w \in V : \{v, w\} \in E(G)\}$, and the closure neighborhood of v is $N[v] = N(v) \cup \{v\}$. The degree of a node v , denoted as $\delta(v)$, is the number of neighbors that it has, that is $\delta(v) = |N(v)|$. A vertex v is *pendant* if its neighborhood contains only one vertex; an edge e is *pendant* if one of its endpoints is a pendant vertex. The degree of the graph G is $\Delta(G) = \max\{\delta(x) : x \in V\}$.

Given a graph $G = (V, E)$, $S = (V', E')$ is a subgraph of G if $V' \subseteq V$ and E' contains edges $\{v, w\} \in E$ such that $v \in V'$ and $w \in V'$. If E' contains every edge $\{v, w\} \in E$ where $v \in V'$ and $w \in V'$ then S is called the *subgraph of G induced by S* and is denoted by $G \parallel S$. We write $G - S$ to denote the graph $G \parallel (V - V')$. In the same way, $G - v$ for $v \in V(G)$ denotes the induced subgraph $G \parallel (V - \{v\})$, and $G - e$ for $e \in E(G)$ is the subgraph of G formed by $V(G)$ and $E(G) - \{e\}$.

A *connected component* of G is a maximal induced subgraph of G , that is, a connected component is not a proper subgraph of any other connected subgraph of G . Notice that, in a connected component, for every pair of its vertices u, v , there is a path from u to v . A tree graph is an acyclic connected graph.

We say that a 2-CF Σ is a *path*, a *cycle*, or a *tree* if its corresponding constrained graph G_Σ is a path, a cycle, or a tree, respectively.¹⁷

3. The Minimal Spanning Tree of a 2-CF

Given a 2-CF Σ , we say that the set of *connected components* of Σ are the subformulas corresponding to the connected components of G_Σ .

Let Σ be a 2-CF. If $\mathcal{F} = \{G_1, \dots, G_r\}$ is a partition of Σ (over the set of clauses appearing in Σ), i.e. $\bigcup_{\rho=1}^r G_\rho = \Sigma$ and $\forall \rho_1, \rho_2 \in [r], [\rho_1 \neq \rho_2 \Rightarrow G_{\rho_1} \cap G_{\rho_2} = \emptyset]$, we say that \mathcal{F} is a *partition in connected components* of Σ if $\mathcal{V} = \{v(G_1), \dots, v(G_r)\}$ is a partition of $v(\Sigma)$.

If $\{G_1, \dots, G_r\}$ is a partition in connected components of Σ , then:

$$\mu_{v(\Sigma)}(\Sigma) = [\mu_{v(G_1)}(G_1)] * \dots * [\mu_{v(G_r)}(G_r)] \quad (1)$$

The different connected components of G_Σ constitute the partition of Σ in its connected components, even if G_Σ is disconnected. In order to compute $\#SAT(\Sigma)$, first we should determine the set of connected components of G_Σ and that can be done in linear time [6]. Then, $\#SAT(\Sigma)$ is reduced to compute $\#SAT(G)$ for each connected component G of G_Σ . From now on, when we mention a 2-CF Σ , we assume that Σ is a connected component graph.

In our case, a *minimal spanning tree* of a connected component G_Σ which corresponds with a 2-CF Σ is a tree, denoted by T_Σ , containing all vertices of G_Σ and such that

1. $\#SAT(A_\Sigma) \geq \#SAT(\Sigma)$
2. $\#SAT(A_\Sigma)$ is minimal into the set of all spanning trees of G_Σ

A *spanning tree collection* for G_Σ is a set of trees, one for each connected component of G , so that each tree is a spanning tree for its connected component. A *minimal spanning tree collection* is a spanning tree collection where each tree has a minimal number of models with respect to any other spanning tree of the connected component.

There are different algorithms for finding a minimal spanning tree in an undirected graph although, as far as we know, all of them work assuming a static weight in each edge. In our case, the edges in G_Σ have not associated a static weight instead they have associated a pair of signs.

If we have a subtree A_Σ which will be extended by one of a possible set of edges $E = \{e_1, e_2, \dots, e_k\}$, each edge with one of its end-points in a node of A_Σ and the other end-point in a node not included in A_Σ . The signs of the edges determine how will be the increase of $\#SAT(A_\Sigma)$ to $\#SAT(A_\Sigma \cup \{e\})$ for just one edge $e \in E$. Then, we have dynamic weights associated to each edge $e \in E$ according to the current configuration of a spanning subtree of G_Σ .

We propose a novel algorithm for building a minimal spanning tree assuming such class of dynamic weights on the edges of the input graph. But before to introduce our proposal, we present some procedures for computing the number of models of a formula for basic topology graphs [1].

4. Linear procedures for #2SAT

For each variable $x \in v(\Sigma)$, Σ a 2-CF, a pair (α_x, β_x) called the *initial charge*, is used for indicating the number of logical values: 'true' and 'false' respectively, that x takes when #SAT(Σ) is being computed.

Procedure A: If Σ is a path:

Let Σ be a path (or a linear chain). Σ can be written (ordering clauses and variables, if it were necessary) as: $\Sigma = \{c_1, \dots, c_m\} = \left\{ \{y_0^{\epsilon_1}, y_1^{\delta_1}\}, \dots, \{y_{m-1}^{\epsilon_m}, y_m^{\delta_m}\} \right\}$, where $\delta_i, \epsilon_i \in \{0, 1\}$, $i \in \llbracket m \rrbracket$ and $|v(c_j) \cap v(c_{j+1})| = 1$, $j \in \llbracket m-1 \rrbracket$. As Σ has m clauses then $|v(\Sigma)| = n = m + 1$.

Let f_i be a family of clauses from Σ built as follows: $f_0 = \emptyset$; $f_i = \{c_j\}_{j \leq i}$, $i \in \llbracket m \rrbracket$. Notice that $f_i \subset f_{i+1}$, $i \in \llbracket m-1 \rrbracket$. Let $SAT(f_i) = \{s : s \text{ satisfies } f_i\}$, $A_i = \{s \in SAT(f_i) : y_i \in s\}$, $B_i = \{s \in SAT(f_i) : \bar{y}_i \in s\}$. Let $\alpha_i = |A_i|$; $\beta_i = |B_i|$ and $\mu_i = |SAT(f_i)| = \alpha_i + \beta_i$.

The first pair (α_0, β_0) is $(1, 1)$ since for any logical value to y_0 , f_0 is satisfied. We compute (α_i, β_i) associated with each variable y_i , $i = 1, \dots, m$, according to the signs: ϵ_i, δ_i of the literals in the clause c_i , by the following recurrence equations:

$$(\alpha_i, \beta_i) = \begin{cases} (\beta_{i-1}, \mu_{i-1}) & \text{if } (\epsilon_i, \delta_i) = (0, 0) \\ (\mu_{i-1}, \beta_{i-1}) & \text{if } (\epsilon_i, \delta_i) = (0, 1) \\ (\alpha_{i-1}, \mu_{i-1}) & \text{if } (\epsilon_i, \delta_i) = (1, 0) \\ (\mu_{i-1}, \alpha_{i-1}) & \text{if } (\epsilon_i, \delta_i) = (1, 1) \end{cases} \quad (2)$$

As $\Sigma = f_m$ then #SAT(Σ) = $\mu_m = \alpha_m + \beta_m$. We denote with $' \rightarrow'$ the application of one of the four rules in the recurrence (2).

Example 1 Let $\Sigma = \{(x_1, x_2), (x_2, \bar{x}_3), (\bar{x}_3, x_4), (x_4, x_5)\}$ be a path, the series (α_i, β_i) , $i \in \llbracket 5 \rrbracket$, is computed according to the signs of each clause, as it is illustrated in the figure (1a). A similar path with 5 nodes but with different signs in the edges is shown in figure (1b).

Notice that, according to figure 1, same signs to the adjacent edges of a same node give a greater value for the number of models than different signs to the adjacent edges at the same node. This principle is the base for building minimal spanning trees, since we are looking for edges which provoke a change of signs when they cross by a node of the graph.

When we count models over a constrained graph, we use *computing threads*. A computing thread is a sequence of pairs (α_i, β_i) , $i = 1, \dots, m$ used for computing the number of models over a path of m nodes.

Procedure B: If Σ is a tree:

Let Σ be a 2-CF where its associated constrained graph G_Σ is a tree. We denote with (α_v, β_v) the pair associated with the node v ($v \in G_\Sigma$). We compute

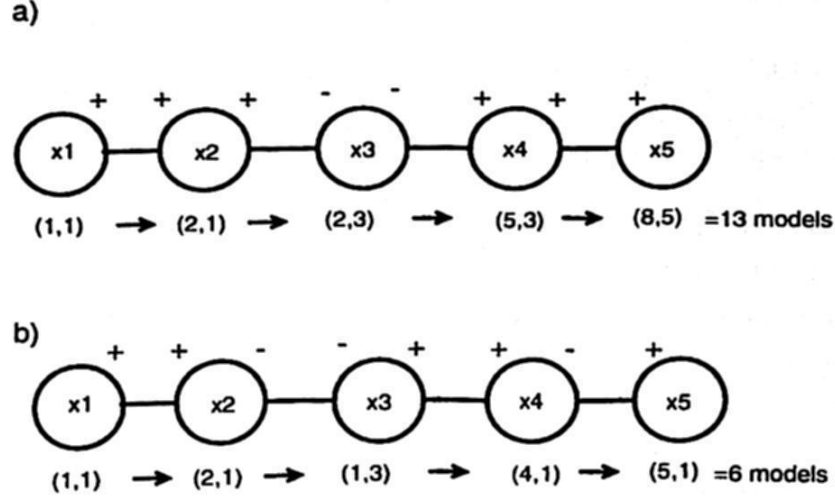


Fig. 1. a) Counting models over paths for monotone formula b) Counting models over paths for non monotone formula

$\#SAT(\Sigma)$ while we are traversing by G_Σ in post-order [7].

Algorithm Count_Models_for_trees(G_Σ)

Input: G_Σ - a tree graph.

Output: The number of models of Σ

Procedure:

Traversing G_Σ in post-order, and when a node $v \in G_\Sigma$ is left, assign:

1. $(\alpha_v, \beta_v) = (1, 1)$ if v is a leaf node in G_Σ .
2. If v is a parent node with a list of child nodes associated, i.e., u_1, u_2, \dots, u_k are the child nodes of v , as we have already visited all child nodes, then each pair $(\alpha_{u_j}, \beta_{u_j})$ $j = 1, \dots, k$ has been determined based on recurrence (2). Then, let $\alpha_v = \prod_{j=1}^k \alpha_{u_j}$ and $\beta_v = \prod_{j=1}^k \beta_{u_j}$. Notice that this step includes the case when v has just one child node.
3. If v is the root node of G_Σ then return $(\alpha_v + \beta_v)$.

This procedure returns the number of models for Σ in time $O(n + m)$ which is the necessary time for traversing G_Σ in post-order.

Example 2 If $\Sigma = \{(x_1, x_2), (x_2, x_3), (x_2, x_4), (x_2, x_5), (x_4, x_6), (x_6, x_7), (x_6, x_8)\}$ is a monotone 2-CF, we consider the post-order search starting in the node x_1 . The number of models at each level of the tree is shown in Figure 2. The procedure *Count_Models_for_trees* returns for $\alpha_{x_1} = 41$, $\beta_{x_1} = 36$ and the total number of models is: $\#SAT(\Sigma) = 41 + 36 = 77$.

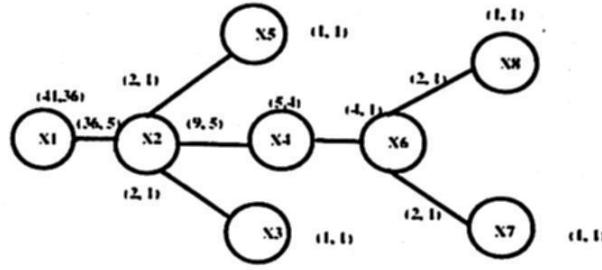


Fig. 2. Counting models over a tree

5. Computation of the Charges of a 2-CF

Once that we know the value $\#SAT(\Sigma)$, the initial charges for all variables of Σ have already been computed.

The final charge (a_x, b_x) of any variable $x \in v(\Sigma)$ is the number of true and false logical values respectively, that x takes into the set of models of Σ , i.e. $\#SAT(\Sigma) = a_x + b_x$. Notice that the initial charge (α_n, β_n) for the last evaluated variable x_n into the above procedures is also its final charge since $\#SAT(\Sigma) = \alpha_n + \beta_n$.

An important result is that we can apply the inverse action realized in each step of the above procedures in order to propagate the final charge to all variables in Σ , in the following way.

Let A_1, \dots, A_n be the sequence of initial charges obtained by the above procedure. Now, we build a new sequence of pairs which represent the final charges (or just the charges) B_n, \dots, B_1 , being B_i the charge of the variable $x_i \in v(\Sigma)$, and which is computed as:

$$\begin{aligned} B_n &= A_n \\ B_{n-i} &= \text{balance}(A_{n-i}, B_{n-i+1}), \quad i = 1, \dots, n-1 \end{aligned} \quad (3)$$

$\text{balance}(A, B)$ is a binary operator between two pairs, e.g. if $x \xrightarrow{s_1 s_2} y$ is an edge of the DAG D_Σ and assuming $A = (\alpha_x, \beta_x)$ be the initial charge of the variable x , $B = (a_y, b_y)$ be the final charge of the variable y , then balance produces a new pair (a_x, b_x) which will be the final charge for x , i.e. $\#SAT(\Sigma) = a_x + b_x$.

Let $\mu_x = \alpha_x + \beta_x$ and $\mu_y = a_y + b_y$. Let $P_1 = \frac{\alpha_x}{\mu_x}$ and $P_0 = \frac{\beta_x}{\mu_x}$ be the proportion of the number of 1's and 0's in the initial charge of the variable x . The charge (a_x, b_x) is computed, as:

$$\begin{aligned} a_x &= a_y \cdot P_1 + b_y; \quad b_x = \mu_y - a_x \quad \text{if}(s_1, s_2) = (+, +) \\ b_x &= b_y \cdot P_0 + a_y; \quad a_x = \mu_y - b_x \quad \text{if}(s_1, s_2) = (-, -) \\ b_x &= b_y \cdot P_0 + a_y; \quad a_x = \mu_y - b_x \quad \text{if}(s_1, s_2) = (+, -) \\ a_x &= a_y \cdot P_1 + b_y; \quad b_x = \mu_y - a_x \quad \text{if}(s_1, s_2) = (-, +) \end{aligned} \quad (4)$$

Note that the essence of the rules in balance consists in applying the inverse operation utilized via recurrence (2) during the computation of $\#SAT(\Sigma)$, and following the inverse order used in the construction of the sequence A_1, \dots, A_n .

Furthermore, in the case of the bifurcation from a father node to a list of child nodes, the application of the recurrence (4) remains valid since each branch has its respective pair of signs.

A special case to consider is when there are two connected components C_1, C_2 where the charges of their variables have been computed, and a new edge $e = \{x, y\}$ with signs (s_1, s_2) will be utilized for joining both components in just one connected component C . Assuming a charge of (α_x, β_x) for x and (α_y, β_y) for y , we must update such charges, indicated by (α'_x, β'_x) for x and (α'_y, β'_y) for y according with the signs in e , in the following way.

$$\begin{aligned}
 \alpha'_x &= \alpha_x * (\alpha_y + \beta_y); \beta'_x = \beta_x * \alpha_y \\
 \alpha'_y &= \alpha_y * (\alpha_x + \beta_x); \beta'_y = \beta_y * \alpha_x \text{ if } (s_1, s_2) = (+, +) \\
 \beta'_x &= \beta_x * (\alpha_y + \beta_y); \alpha'_x = \alpha_x * \beta_y \\
 \beta'_y &= \beta_y * (\alpha_x + \beta_x); \alpha'_y = \alpha_y * \beta_x \text{ if } (s_1, s_2) = (-, -) \\
 \alpha'_x &= \alpha_x * (\alpha_y + \beta_y); \beta'_x = \beta_x * \alpha_y \\
 \beta'_y &= \beta_y * (\alpha_x + \beta_x); \alpha'_y = \alpha_y * \beta_x \text{ if } (s_1, s_2) = (+, -) \\
 \beta'_x &= \beta_x * (\alpha_y + \beta_y); \alpha'_x = \alpha_x * \beta_y \\
 \alpha'_y &= \alpha_y * (\alpha_x + \beta_x); \beta'_y = \beta_y * \alpha_x \text{ if } (s_1, s_2) = (-, +)
 \end{aligned} \tag{5}$$

Notice that if some of the initial charges are (1,1) the above recurrence is equivalent with its corresponding case (by the signs) in equation (4).

After computing the new charges for x and y , the charges for all the remaining variables in C have to be updated, for propagating the new values (α'_x, β'_x) to the original variables of C_1 and (α'_y, β'_y) to the original variables of C_2 according to the operator *balance*.

Notice that the computation of the charges of a 2-CF Σ has the same complexity order that the one used for computing $\#SAT(\Sigma)$. Thus, if the constrained graph of Σ does not contain cycles, then we compute all the charges of the variables of Σ in polynomial time [1].

6. Building the Spanning tree of the Constrained Graph

Let $G_\Sigma = (V, E, \{+, -\})$ be a signed connected graph of an input formula Σ in 2-CF.

Let v_r be a node of G_Σ chosen to start a depth-first search. Each back edge $c_i \in E$ found during the depth-first search marks the beginning and the end of a fundamental cycle of G_Σ .

If the input formula Σ is in fact a tree then the output of the algorithm is the same tree and we just apply the procedure (c) for computing the number of models: $\#SAT(\Sigma)$.

In the case when there are cycles in G_Σ then we apply the following procedure in order to determine a minimal spanning tree A_F of G_F .

Our proposal works like the well known Kruskal's algorithm. An initial spanning tree $A_\Sigma = (V(G_\Sigma), P_Edges)$ is formed by all vertices of G_Σ since all vertices are connected components by themselves, and all pendant edge of G are

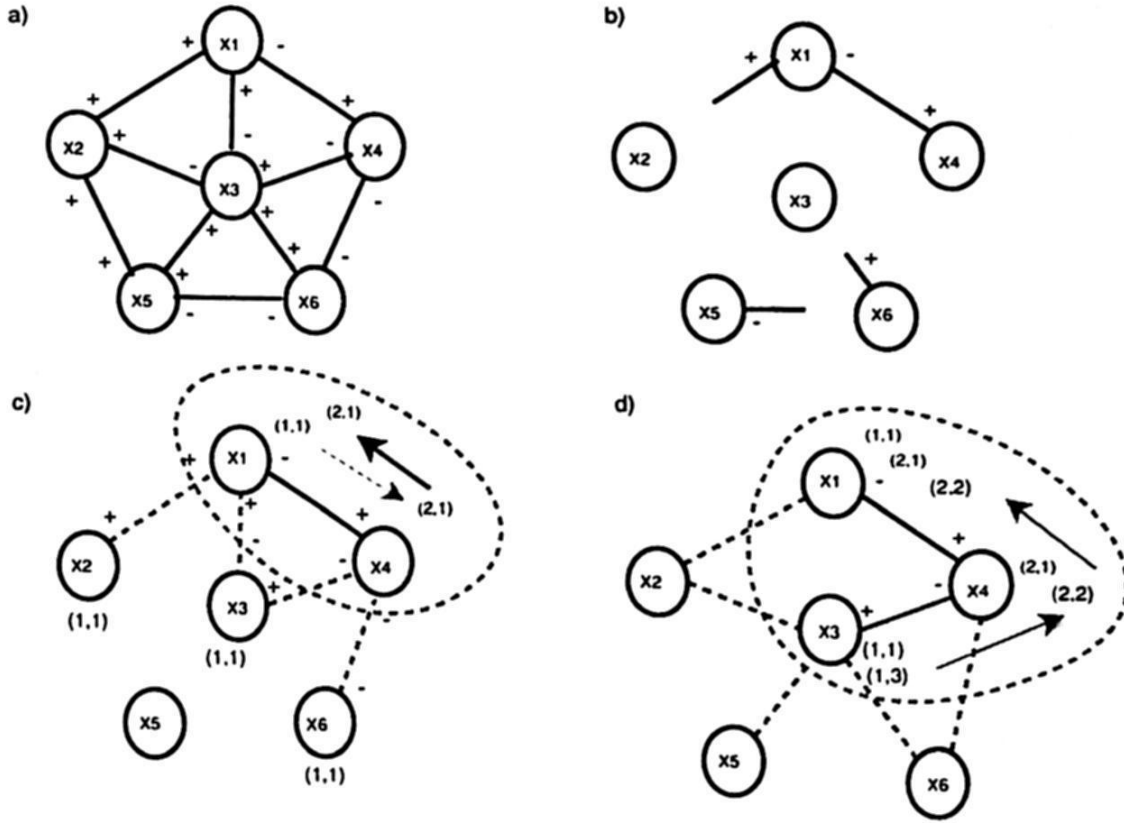


Fig. 3. a) Original Graph b) Selecting of an edge candidate c) Building the first connected component with its respective setback d) Adding a node to the component

edges of the spanning tree (if there are not pendant edges then an emptyset is initially assigned to A_{Σ}).

In each step of the algorithm *Spanning_Tree*, the procedure *Count_Models* reviews the increment on the number of models when an edge $e \in (E(G_{\Sigma}) - E(A_{\Sigma}))$ is considered for being added to the spanning tree.

In order to extend the connected components in A_{Σ} , the edges in $(E(G_{\Sigma}) - E(A_{\Sigma}))$ which conform cycles with A_{Σ} are deleted from $(E(G_{\Sigma}))$. And the remaining edges are ordered according to the increment on the number of models in $\#SAT(A_{\Sigma})$. If $e \in (E(G_{\Sigma}))$ infers a minimal increment on $\#SAT(A_{\Sigma})$ with respect to the any other edge, e is selected to be added to A_{Σ} . Notice that the increment on the number of models depends mainly of the signs associated to e as well as the charge of the two-endpoints of e .

There are a set of strategies for detecting the edges in $(E(G_{\Sigma}) - E(A_{\Sigma}))$ which infer a minimal increment on the number of models in the spanning tree A_{Σ} and in fact, when the remaining edges in G have similar values of increment on the number of models such strategies are also applied. Such strategies are:

Algorithm 1 Procedure *Spanning_Tree*(G_Σ)

Input: $G_\Sigma = (V(G), E(G))$ {a constrained signed graph}
 Initiate:
 Let $P_Edges = \{e \in E(G_\Sigma) : e \text{ is a pendant edge}\}$;
 $All_Edges := E(G) - P_Edges$; {Set of initial edges to test}
 $A_\Sigma := (V(G), P_Edges)$; {all node and pendant edge are connected components of the Tree}
 $Cs := \emptyset$; {Set of potential edges which make a change of sign on some vertices}
 Iter := 1; {The first iteration}
while ($All_Edges \neq \emptyset$) **do**
 $Count_Models(All_Edges, Vect_Models)$; {count the new number of models generated by each potential edge}
 if ($Vect_Models_has_different_values$) **then**
 $Sel_Edge = \min\{Vect_Models\}$; {select the edge which increases a minimum the number of models}
 else
 $Cs = Find(Test, A_\Sigma)$; {looking for edges which could generate a change of sign in any node of A_Σ }
 $Test = complete(Cs)$; {choose edges where its two end-points generate a change of sign on the nodes}
 $Sel_Edge = First(Test)$; {Select the edge with keeps a potential change of sign of a node}
 end if
 $All_Edges := All_Edges - \{Sel_Edge\}$;
 $E(A_\Sigma) := E(A_\Sigma) \cup \{Sel_Edge\}$;
 $All_Edges := All_Edges - Edges_Cycles(A_\Sigma, All_Edges)$; {delete all edge which conform a cycle with the tree A_Σ }
end while

1. If e connects two different connected components of A_Σ where its endpoints v_i and v_j have a change of sign over its incident edges then e is an optimal selection.
2. In general, if two edges e_1 and e_2 generate the same increment on the number of models of A_Σ e_1 is preferred over e_2 if e_1 could bring about a change of signs, in the following steps, on its incident node.
3. When two connected components are joining for forming just one, it is preferable to obtain a path over a tree, as a resulting new connected component.

Thus, we build in polynomial time a spanning tree A_Σ from G_Σ such that

- $\#SAT(A_\Sigma) \geq \#SAT(\Sigma)$
- $\#SAT(A_\Sigma)$ is minimal into the set of all spanning tree of G_Σ

The application of Algorithm 1 to a graph is shown in Figure 3, and finally the minimal spanning tree is shown in Figure 4.

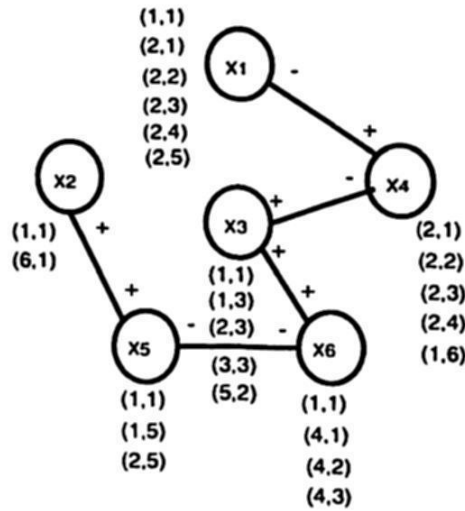


Fig. 4. A Minimal Spanning Tree resulting from the application of the algorithm 1, its number of models is $(6,1)=7$

7. Conclusions

Although the #2SAT Problem is a #P-complete problem, given a 2-CF Σ there are different methods for computing in an approximate way the value $\#2SAT(\Sigma)$. One of those methods, is based on computing the minimal spanning tree A_{Σ} .

Although the edges of the constrained graph have not weights, there are a pair of signs associated with each edge. And in this case, the signed edges allow us to determine dynamic weights which are the base for computing a minimal spanning tree of the constrained graph of a Boolean two conjunctive form.

References

1. De Ita G., Bello P., Contreras M., Efficient counting of models for Boolean Formulas Represented by Embedded Cycles, *CEUR WS Proceedings*, 286, (2008).
2. Dyer M., Greenhill C., Some #P-completeness Proofs for Colourings and Independent Sets, Research Report Series, University of Leeds, 1997.
3. Garey M., Johnson D., Computers and Intractability a Guide to the Theory of NP-Completeness, W.H. Freeman and Co., 1979.
4. Greenhill Catherine, The complexity of counting colourings and independent sets in sparse graphs and hypergraphs", *Computational Complexity*, 9(1): 52-72, 2000.
5. Russ B., *Randomized Algorithms: Approximation, Generation, and Counting*, Distinguished dissertations Springer, 2001.
6. Roth D., On the hardness of approximate reasoning, *Artificial Intelligence* 82, (1996), pp. 273-302.
7. Tarjan R., Depth-First Search and Linear Graph Algorithms, *SIAM Journal on Computing*, Vol. 1, pp.146-160, 1972.
8. Vadhan Salil P., The Complexity of Counting in Sparse, Regular, and Planar Graphs, *SIAM Journal on Computing*, Vol. 31, No.2, pp. 398-427, 2001.

A Hybrid Evolutionary Algorithm for the Edge Crossing Minimization Problem in Graph Drawing

Sergio Enriquez¹, Eunice Ponce de León¹, Elva Díaz¹ and Alejandro Padilla¹

¹ Departamento de Ciencias de la Computación,
Centro de Ciencias Básicas,
Universidad Autónoma de Aguascalientes
(senriquez, eponce, ediazd, apadilla)@correo.uaa.mx

Abstract. This document explains the design and implementation of a hybrid evolutionary algorithm for the edge crossing minimization problem in graph drawing. This algorithm, called HUX, combines a global search algorithm (EDA — Estimation of Distribution Algorithm) with a local search Algorithm (HC — Hill Climbing) in order to establish a balance between the exploration and exploitation efforts. The HUX has shown to be more efficient and robust in search of the optimum solution in comparison to other meta-heuristics, such as HC, Univariate Marginal Distribution Algorithm and the Genetic Algorithm, which was also implemented in this comparison. Experiments were performed using planar and non-planar graphs. The quality and frequency of the optimum solutions registered by the hybrid HUX algorithm were higher than those registered by other algorithms.

Keywords: Graph Drawing, Crossing Minimization, Hill Climbing, Estimation of Distribution Algorithms, Genetic Algorithms.

1 Introduction

The problem in finding the minimum number of crossing edges in a graph entails finding, within all possible forms of drawing a graph, the one that includes the minimum number of crossings on its edges. The problem is difficult because as Garey and Johnson [7] showed is NP-hard. One of the main problems involved in the clear visualization of a graph is the problem of the cross minimization of the graph's edges [12]. This is a typical and very important problem that arises in graph drawing. Readability is reduced in graphs with a large number of crossings, especially in the case of dense graphs. The graph visualization problem is currently open to research; it can be applied to different areas, as in biology and chemistry, object oriented systems, data structures, real time systems, flowcharts, entity relation charts, semantic nets, project management, representation of knowledge charts, logical programming, design of VLSI circuits, virtual reality, cartography, and social networks, among others [8].

There are different Evolutionary Algorithms (EAs) which deal with the edge crossing minimization problem in graph drawing; however, most of these are based on genetic algorithms. The Estimation of Distribution Algorithms (EDAs) [10] has been poorly used in Graph Drawing (GD), for example, the only EDA algorithm known by the authors is UMDA in [14].

In this paper a hybrid algorithm composed with a global search algorithm (UMDA—Univariate Marginal Estimation Algorithm) and a local search algorithm (HC—Hill Climbing) is introduced. The HUX (Hill Climbing and UMDA x-crossing) has proven more efficient with regard to optimum solution search (in this case, the drawing which includes the least number of crossings on the graph's edges) than metaheuristics HC, UMDA and the Simple Genetic Algorithm (SGA) which was also implemented in this comparison. Measurement of results was carried out using different planar (with no crossings among its edges) and non-planar graphs (with at least one crossing among its edges), which were the measurement subjects to which the various already mentioned metaheuristics were applied. The quality and frequency of optimum solutions registered by the hybrid HUX algorithm were higher than those registered by the algorithms against which this measurement was carried out and that forms part of the state of the art within graph drawing. It must be mentioned that in each case studied, the hybrid HUX metaheuristic always finds the optimum for benchmarks and, additionally, the occurrence frequency of the optimum was 15% greater than the second best algorithm of all those studied. A proper graphical interface called Minx (Minimization x-crossing) was developed for graph visualization. This interface shows the automatic drawing of each graph obtained by the metaheuristics implemented in this paper.

2 Contents

The contents in this document are intended to:

- (1) Define the optimization problem of the edge crossing minimization problem in graph drawing (section 3)
- (2) Define hybrid base algorithms: UMDA algorithm and HC algorithms and their pseudo-code (section 4)
- (3) Define the implemented hybrid HUX algorithm (section 5)
- (4) Define how experiment designs were implemented (section 6)
- (5) Compare results among the various implemented algorithms, and to compare results with the hybrid HUX algorithm results (section 7)
- (6) Discuss and draw conclusions with regard to main research's contributions (section 8)

3 Optimization problem, solution representation and evaluation function

The graph edges crossing number optimization problem can be described as follows [5]:

Let $G = (V, E)$ be a graph, let V be a set of vertices and E a set of edges. MG is the adjacency matrix and P the Cartesian plane. The problem involves finding a pair $(x, y) \in P$ for each $v \in V$ such that the number of crossings between the edges of the graph is minimum. Each pair (x, y) represents a position of the vertex v in Cartesian Plane P . For any two vertices $v = (x, y)$ $v' = (x', y')$, $x \neq x'$ and $y \neq y'$. In this paper, the graph's edges are considered straight lines.

Let N be the number of vertices of the graph. The solution representation S (i.e., a graph draw, or graph layout) is as follows:

$$S = (x_1, y_1, x_2, y_2, \dots, x_i, y_i, \dots, x_N, y_N)$$

Each pair x_i, y_i represents the position of the i -th vertex in Cartesian Plane P .

A graph draw S is evaluated by number of crossings of graph edges in this graph draw S . The number of crossings was obtained by solving an equation system for all pair of edges of the graph. The following cases are analyzed: crossing edges, overlapping edges, and overlapping vertices.

4 Base Algorithms for Hybridization

This section gives an explanation of the algorithms used for carrying out the hybridization of the proposed algorithm.

4.1 Univariate Marginal Estimation Algorithm (UMDA)

Introduced by Mühlenbein, [11] this is a particular case of EDAs which is considered as having no dependencies; its distribution of n -dimensional joint probability factorizes as a product of n univariate and independent probability distributions.

Example:

$$p_i(x) = \prod_{i=1}^n p_i(x_i) \quad (1)$$

The joint probability distribution of each generation was estimated from individuals $p_i(x)$ selected. The joint probability distribution factorizes as the product of independent univariate distributions.

Example

$$p_i(x) = p(x | D_{i-1}^{Sc}) = \prod_{i=1}^n p_i(x_i) \quad (2)$$

Every univariate probability distribution is estimated by marginal frequencies:

$$p_i(x) = \frac{\sum_{j=1}^N d_j(X_i = x_i | D_{i-1}^{Se})}{N}$$

where:

$$d_j(X_i = x_i | D_{i-1}^{Se}) = \begin{cases} 1 & \text{if on the } j\text{-th case of } D_{i-1}^{Se}, X_i = x_i \\ 0 & \text{in another case} \end{cases} \quad (3)$$

Pseudocode UMDA

D_0 Generate M individuals at random (initial population)
Repeat for $l = 1, 2, \dots$ until stop criterion is verified.

$D_{l-1}^{Se} \leftarrow$ Select $N \leq M$ individuals from D_{l-1} according to selection method.

$$p_i(x) = p(x | D_{l-1}^{Se}) = \prod_{i=1}^n p_i(x_i) = \prod_{i=1}^n \frac{\sum_{j=1}^N d_j(X_i = x_i | D_{l-1}^{Se})}{N} \leftarrow \begin{array}{l} \text{Obtain} \\ \text{estimate of} \\ \text{combined} \end{array}$$

probability distribution D_l Sample M individuals (new population) from $p_l(x)$.

4.2 Hill Climbing Algorithm

The hill climbing algorithm (HC) [15] is an optimization technique which belongs to the local search family. This algorithm uses a series of iterations where it is constantly shifting towards the direction with a better value. When the algorithm reaches a point where its result cannot be further improved, it needs to start all over at another point where it can direct its search. This is achieved by a random restart of the algorithm. This technique performs a series of climbing the top searches from randomly generated initial states. The best result obtained thus far is saved as the algorithm's iterations progress and it stops when no significant progress has been accomplished. The algorithm's stop condition may be set by a fixed number of iterations or when the best result has not improved over the course of a number of iterations.

Pseudocode HC

function HILL_CLIMBING(problem) returns a solution
state

inputs: problem, a problem

static: current, a node

next, a node

current ? MAKE-NODE(INITIAL-STATE[problem])

loop do

```

    next ? a highest-valued successor of current
    if VALUE[next]  $\leq$  VALUE[current] then return
    current ? neighbor
end

```

5 Proposed hybrid HUX Algorithm

The hybrid HUX algorithm [6] [4] is based mainly on the UMDA algorithm; hence, it was necessary to modify this algorithm. Hybridization of the algorithm is achieved by applying 10 cycles of the HC algorithm with random restart to the best individual in the population. The HC algorithm was implemented by means of elitism because the algorithm was held back at a local optimum.

Hybridization performed in the UMDA algorithm took place as follows:

- Elitism was applied only if the best individual in the population had a better degree of adaptability to the problem (fewer crossings) as compared to the worst individual in the population. In this case, the best individual in the population is saved in the worst individual, thus rearranging the whole population.
- Hybridization was performed by applying 10 cycles of the HC with random restart to the best individual in the population. The result of this hybridization has the potential of improving the individual's degree of adaptability (which is what happens in most cases) or, it is possible for the degree of adaptability to worsen. Were this to be the case, convergence of the algorithm is achieved regardless because at the time of implementing elitism in the population, the best solution found thus far is never lost.

5.1 Pseudocode HUX

```

// for hybridization and elitism
public class UMDA {boolean HUX;
    public UMDA(boolean HUX) {/* constructor for
                                hybridization */
        this.HUX = HUX;
    }
    PopG oldpop, newpop;
    IndividualG best;
    //generate initial population
    For (i = 0; i < size_pob; i++)
        oldpop[i] = Individual;
    //random generation of vertices
    For (j = 0; j < total_vertex; j++)
        oldpop[i].vertex[j] = generateVertex;
    newpop = oldpop; //algorithm cycle
    //algorithm cycle for new population

```

```

For (i = 0; i < max_generation; i++)
{
    evaluateFobject;      /* evaluates population
                           quality */
    sortPop;              /* descending sort by total
                           cross */
    //initiates implementation of hybridization
    if this.HUx then
    {
        elitism; //elitism applied the population
        applyHC; /* 10 cycles of the HC applied to
best individuals*/
        // end of implementation hybridization
        best = saveBest;  /*save the best individual
        calculaVectoProb; /* estimated
                           probability vector */
        oldpop = newpop;  /* copy new population
                           in older population */
    }
}

```

6 Experiment Design

In order to compare the four algorithm implemented in this paper, seven graphs were used to perform experiments. The seven graphs were selected from the papers [14],[1], [3], [9], [13] to use them as benchmarks. Table 1 shows the general aspects of all graphs researched. This table shows the total number of vertices and edges in each graph as well as a density column, which is the ratio between existing edges in the graph and the maximum possible number of edges, which would be the complete graph K_p , having the same number of vertices p [14].

Table 1. General properties of graphs taken from literature.

Graph	Vertices (p)	Edges (q)	Density
Simple	8	12	0.429
Herschel	11	18	0.327
Hobbs	20	36	0.189
Tree	34	33	0.059
Star	17	40	0.294
Grid 4 x 4	16	24	0.200
Composite	40	69	0.088

The Simple graph was taken from [13] and it is the simplest of all graphs, having 8 vertices and 12 edges. The Herschel graph, taken from [1] has 11 vertices and 18 edges. The tree graph, taken from same paper, has 34 vertices and 33 edges and Star

graph, also taken from the same paper has 17 vertices and 40 edges. The Hobbs graph was taken from the [9] and it has 20 vertices and 36 edges. The characteristic of this graph is that it is non-planar. The Grid graph was taken from [3]. This graph has 16 vertices and 24 edges and it corresponds to a 4 x 4 grid. The composite graph is the largest of all graphs, having a total of 40 vertices and 69 edges. This is a proper graph, which was created to test the hybrid HUX algorithm's stability before graphs with a greater number vertices and edges as described above. The composite graph [4] was built using vertices and edges from the Hobbs graph [9] as well as the Ebner graph [2]. The Ebner graph has a total of 20 vertices and 33 edges. This graph was taken from the literature and it was used only for building the composite graph.

All algorithms implemented in this paper were executed using the same number of iterations for all graphs taken from the literature, 3,600 evaluations in the case of the simple graph and 20,000 evaluations in all other instances. Additionally, 20 executions were performed independently for every individual algorithm. The average number of crossings found by the metaheuristics which were implemented demonstrates that the hybrid HUX algorithm outperformed all other algorithms.

The GA uses the following parameters; the reproduction cycle included 500 generations, a population composed by 40 individuals, a 90% crossover probability, and a 10% mutation probability. The above parameters permit the accomplishment of best results for this particular algorithm.

The algorithm HC with random restart also employs iterations (cycles) when searching for the best solution; therefore, we worked with 20,000 iterations. In this algorithm, one of the graph's vertices is selected at random and its coordinates are replaced by new vertex coordinates which are randomly generated in the plane.

The UMDA used 200 generations along with a population size containing 100 individuals, and a 50% truncation rate of the population. These parameters were the ones that yielded the best results for this particular algorithm.

The HUX algorithm uses a method of elitism in order to avoid losing the best solution found. It also uses 10 cycles of the up hill climber with random restart. This algorithm's parameters are as follows: a total of 200 generations, a population size containing 100 individuals, and a 50% truncation rate of the population. This is the top combination of parameters that yield the best results for this particular algorithm.

In the case of Composite graph, 800 generations were used, along with a population size containing 100 individuals, a 90% crossover probability and 10% mutation probability for population-based algorithms (GA, UMDA and HUX). In the case of the HC algorithm, 80,000 evaluations were performed (800 generations x 100 individuals) for finding the best solution. These parameters are the ones that yielded the best results for each algorithm.

In general, we used a total of 3,600, 20,000 and 80,000 evaluations (cycles) and a total number of 20 independent executions for each algorithm. This measurement intends to compare results obtained by our algorithms to those obtained from a group of papers included in the literature, in which graphs are drawn by means of general purpose methods, such as the Genetic Algorithm, the Hill Climbing algorithm and Univariate Marginal Estimation Algorithm. For instance, the graph designated as Simple [14] used 3,600 evaluations per execution. In order to compare it with the algorithms used in this paper, it was necessary to carry out 60 generations on 60 individuals for the population-based algorithms (GA, UMDA and HUX). This section

lists all algorithms employed for completing the hybridization of the proposed algorithm.

7 Results and Discussion

The results of this experiment appear on tables 2, 3, 4, 5, and the discussions as follows.

7.1 Comparison of results among algorithms

Table 2 shows the overall average number of crossings found by each algorithm, as well as the overall average number of crossings found for each graph taken from the literature. This table shows how the hybrid HUX algorithm averages, for the most part, slightly less than 1 crossing per graph for the 20 executions carried out for each one, followed by the HC with random restart. Likewise, it can be observed that the hybrid HUX algorithm is the best of all algorithms with an overall average of 1.03 crossings per graph, while the worst one is the GA algorithm, whose overall average is 5.87 crossings per graph.

Table 2. Average number of crossings for graphs obtained from the literature.

Graph	Average number of crossings found				Total
	GA	UMDA	HUX	HC (random reboot)	
Simple	0.00	0.00	0.00	0.00	0.00
Herschel	1.70	2.05	0.40	0.75	1.23
Tree	7.70	1.05	0.25	1.20	2.55
Hobbs	11.35	9.90	4.80	6.15	8.05
Star	12.60	8.20	0.55	0.70	5.51
Grid	1.65	1.10	0.15	0.80	0.93
Total	5.83	3.72	1.03	1.60	

Further analyzing the results, one can see that the Hobbs graph [9] is the most difficult graph, having an overall average of 8.88 crossings found for all the algorithms implemented, while the Simple graph is the easiest of all, having an overall average of 0 crossings found for every algorithm implemented in this paper.

Table 3 shows the optimal frequency of occurrence reported in each of the graphs implemented in this paper. The frequency is defined by the number of times the algorithm reached the optimum. The table shows the optimal frequency of occurrence reached in the 20 independent evaluations that were executed for each algorithm.

Table 3. Number of times the algorithm found the optimum in a total of 20 evaluations.

Graph	Optimal frequency of occurrence			
	GA	UMDA	HUx	HC (random reboot)
Simple	20	20	20	20
Herschel	8	2	16	14
Tree	0	7	15	8
Hoobs	0	0	6	3
Star	0	2	17	16
Grid	6	9	17	12
	28%	33%	76%	61%

Table 4 shows the optimal frequency of occurrence. The HUx algorithm outperforms all other algorithms, in some instances by more than 50%. The following table shows the percentage of occurrence of the optimum and it can be seen that the HUx algorithm has an effectiveness rate of 76% in obtained results, versus the 61% rate shown by the HC algorithm with random reboots. This is a 15% difference with regard to the effectiveness rate as compared to the second best algorithm.

Table 4. Average optimal frequency of occurrence in a total of 20 evaluations.

Graph	Average Frequency of occurrence of the optimum			
	GA	UMDA	HUx	HC (random reboot)
Simple	100%	100%	100%	100%
Herschel	40%	10%	80%	70%
Tree	0%	35%	75%	40%
Hoobs	0%	0%	30%	15%
Star	0%	10%	85%	80%
Grid	30%	45%	85%	60%
	28%	33%	76%	61%

Table 4 shows that the percentage of occurrence of the optimal HUx algorithm lies within an effectiveness rate of 100% and 75% in the case of the Simple graph, the Herschel graph, the Tree graph, the Star graph and Grid graph. Only the Hobbs graph, which is the most difficult of all, a 30% effectiveness rate was obtained. Nonetheless, it still exceeded the 50% effectiveness rate shown by the HC algorithm with random reboot, which is the second best of all algorithms.

Finally, it is worth mentioning that the reach of our objective, which in this case entailed a drawing having the minimum number of crossings in the graph's edges, was fully achieved by the hybrid HUx algorithm as well as by the HC algorithm with random reboot. The search for the optimal solution for all graphs taken from the literature includes zero crossings on the graph's edges; only in the Hobbs graph which is non-planar, in this case the optimal solution is 2 crossings on the edges of this graph.

Table 5 shows the reach of goals accomplished by the different metaheuristics for each of the graphs taken from the literature.

Tabla 5. Reach of the goals obtained by the different metaheuristics.

Graph	Scope of Objectives			HC (random reboot)
	GA	UMDA	HUx	
Simple	0	0	0	0
Herschel	0	0	0	0
Tree	2	0	0	0
Hoobs	5	3	2	2
Star	2	0	0	0
Grid	0	0	0	0

When analyzing the results obtained by the UMDA algorithm and the HC algorithm with random reboot, it is obvious that the reach of their objectives is very similar, taking into account that this feature makes the decision to implement a new algorithm having the virtues of these two algorithms in order to overcome the results possessed these two algorithms up to that moment. This is how the hybrid HUx algorithm is implemented.

7.2 Comparison of HUx algorithm with literature results

A comparison of the results obtained by the hybrid HUx algorithm and some of the algorithms reported in the literature comprising part of the state of the art on graph drawing for the minimization of crossings on their edges is shown below. The way results are presented in this section differs from the way results were presented in the previous index because most of the papers herein cited does not allow a direct comparison given the fact that only in a few instances are statistical summaries provided. Hence, the authors only give a description and show some of the results obtained. Therefore, the way of comparing the results for each graph is done separately.

The Simple graph for the paper presented in [13], yields 93% of drawings with no crossing after 3,600 evaluations (60 generations x 60 individuals). With regard to the same graph presented by [14], the Stochastic Hill Climbing (SHC), yields a 100% drawings with no crossings after 550 evaluations, while the hybrid HUx algorithm yields 100% of drawings with no crossings after 252 evaluations on average in a total of 20 independent executions. This is a better result since it uses only half of the evaluations to achieve the goal with respect to the SHC.

The paper of [99] yields a solution with four crossings after 20,000 evaluations (1000 generations x 20 individuals). The SHC [14], yields two solutions with three crossings, four solutions with four crossings and three solutions with five crossings during the 20 executions, and no details are given with regard to number of evaluations needed to accomplish the solution. The hybrid HUx algorithm presents six solutions with two crossings, three solutions with three crossings, two solutions with four crossings and two solutions with five crossings after 9,034 evaluations on

average during the 20 executions of the algorithm. These results are by far superior to the those obtained by Hobbs and Rossete, since six solutions were obtained with two crossings, something not achieved by the SHC algorithm.

The results presented by [1] for the tree graph yields only one solution with no crossings. The SHC [14], yielded ten solutions with no crossings and six solutions with one crossing after 3,000 evaluations. The hybrid HUX algorithm yielded fifteen solutions with zero crossings and five solutions with one crossing in just 4,841 evaluations on average in the twenty executions. These results also favor the hybrid HUX algorithm since it yielded five more solutions with zero crossings than the SHC algorithm.

For the Star graph [1], authors obtained one solution with no crossings. The SHC [14] yielded five solutions with no crossings and two solutions with one cross during the 20 executions, no details are given with regard to the number of evaluations required to attain these results. The hybrid HUX algorithm yielded seventeen solutions with zero crossings, one solution with one crossing, one solution with three crossings and one solution with seven crossings upon completing 4909 evaluations on average during the 20 executions for the algorithm. The results obtained for this graph are also much higher since the HUX yielded twelve results with zero crossings more than SHC algorithm.

The Herschel graph SHC [14], yielded seventeen solutions with no crossings after 3,000 evaluations during the 20 executions. The authors of the paper [1] only obtained one solution with no crossings when using the mutation based on the springs algorithm. However, when it is not used, it is not possible to obtain one single solution with no crossings. The hybrid HUX algorithm yielded 16 solutions with no crossings after 4,194 evaluations on average and four solutions with two crossings after 20 executions of the algorithm. For this graph, results obtained by the SHC and hybrid HUX algorithm are very similar, although the SHC obtained included only 1 more crossing than the hybrid HUX algorithm.

The Grid graph [3], yields a solution with no crossings upon completion of 50,000 evaluations (5000 generations x 10 individuals). In the paper of [14], the SHC yielded 12 solutions with no crossings after 3,000 evaluations during the 20 executions of the algorithm, while the hybrid HUX algorithm yielded seventeen solutions with zero crossings and three solutions with one crossing after 4,018 evaluations on average during the 20 executions of the algorithm. The results again favored the hybrid HUX algorithm since it yielded five crossings more than the SHC.

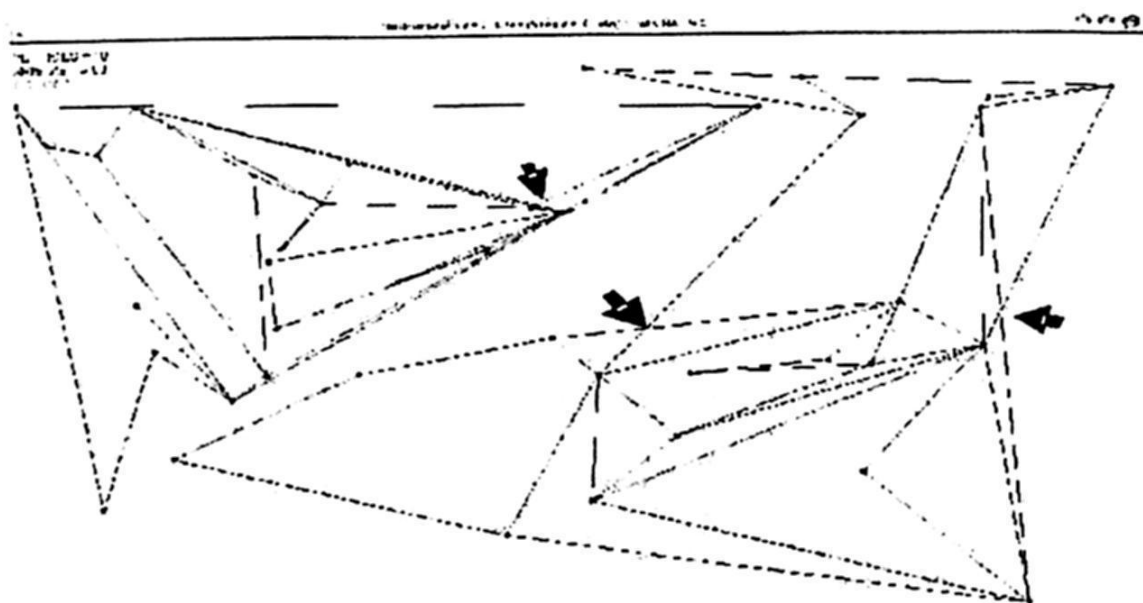
A larger graph was designed with more vertices and edges than graphs taken from the literature. This graph include the vertices and edges of two graphs; namely, the Hobbs graph which has 20 vertices and 36 edges, and the Ebner graph, which has 20 vertices and 33 edges. The graph resulting from the combination of these two graphs was named "Composite graph". This union yielded the Composite graph, which was built having 40 vertices and 69 edges, which is the largest graph of all. The results of the composite graph are shown below in Table 6.

Table 6. Average optimal frequency of occurrence in a total of 20 evaluations.

Algorithm	Composite graph			Average frequency of occurrence
	Average Crossings	Objective Scope least 3 crossings	Frequency of occurrence	
GA	63	35	0	0%
UMDA	34	19	0	0%
HUx	11	3	1	5%
HC (reboot)	10	5	0	0%
	29.542	15.50	0.25	1%

The table's first column shows that the average number of crossings is 11 for the hybrid HUx algorithm, while the HC algorithm with restart obtained a total of 10 crossings, which is one less than the hybrid HUx algorithm, making it the best of all algorithms; however, this result is very similar to that obtained by the hybrid HUx algorithm. Furthermore, it is known that the minimum number of crossings of the Composite graph is 3 because the minimum number of crossings for the Ebner graph is 1, and the minimum number of crossings for the Hobbs graph is 2, as shown in Table 6. By looking at column two on the Table 6, it can be observed that the only algorithm that achieved the objective is the hybrid HUx algorithm, as shown by columns three and four on Table 6. Here one can observe that the only algorithm that contributing to the statistical data is the hybrid HUx algorithm.

Figure 1 shows the Composite graph drawing, which was minimized by the hybrid HUx algorithm. The HUx was the only algorithm that reached the three crossings objective. The Minx System displays the Composite graph [4].

**Fig. 1** Composite graph displayed by the system Minx.

8 Conclusions

Among the main contributions generated by this research paper, the creation of the hybrid HUX algorithm that is used in the graph drawing crossing minimization of edges problem efficiently is included. This can be used in automatic graph drawing.

It was demonstrated that this tool was implemented successfully in all experiments carried out; therefore, it is an efficient way to solve the problem of minimizing the crossings of the edges of a graph. Furthermore, analysis results showed that this tool exceeded the quality of results produced by the various metaheuristics implemented in this research paper as well as the results obtained by other metaheuristics taken from the literature and that make up part of the state of the art.

The design, development and deployment of the hybrid HUX metaheuristic, has been of paramount importance as it leverages the most outstanding properties of two metaheuristics such as the UMDA which is a global search metaheuristic and the HC algorithm is an algorithm of local search. UMDA characteristics to capture global properties of the solutions by estimating a probabilistic model (independence model) and secondly the speed and efficiency shown by the HC algorithm makes the hybrid HUX algorithm a tool most effective in the search for minimization of crossings on the edges of a graph.

The Minx system reported in [4] provides users a tool for a friendly and easy to use graphs display. The automatic drawing of minimized graphs makes it easier for the user to compare results appearing in separate windows, giving the user the opportunity to choose the graph design which best suits their needs.

References

1. Branke, J., Bucher, F., Schmeck, H.: A genetic Algorithm for drawing undirected graphs. *Proceedings of 3rd Nordic Workshop on Genetics Algorithms and their Applications*, Alander, J. T. (Ed.), pp. 193-206 (1997).
2. Ebner, D.: *Optimal Crossing Minimization Using Integer Linear Programming*, Unity Technology of Vienna, 87 pages (2005).
3. Eleoranta T., Mäkinen E.: *A genetic Algorithm for Drawing Undirected Graphs*, Department of Computer and Information Science, University of Tampere Finland, Work supported by the Academy of Finland (Project 35025) (2001).
4. Enriquez S.: *Metaheurística Evolutiva Híbrida para la Minimización de Cruces de las Aristas en el Dibujo de Grafos*, Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Tesis y Disertaciones Académicas, page 91 (2009).
5. Enriquez S., Ponce de León E., Díaz E.: Calibración de un Algoritmo Genético para el Problema de la Minimización de Cruces en las Aristas de un Grafo, *Avances en Computación Evolutiva*, Memorias del IV Congreso Mexicano de Computación Evolutiva, Centro de Investigación en Matemáticas (CIMAT), pp. 61-66 (2008).
6. Enriquez S., Ponce de León E., Díaz E., Padilla A.: Sistema Minx para el Dibujo de Grafos con el Menor Número de Cruces en sus Aristas, *Memorias del Cuarto Congreso Estatal la Investigación en el Posgrado*, Universidad autónoma de Aguascalientes, <http://posgrado.uaa.mx/posgrado/>, page 3 (2008).
7. Garey M. R., Johnson D. S.: Crossing number is NP-complete. *SIAM Journal on Algebraic and Discrete Methods*, vol. 4, pp. 312-316 (1983).

8. Herman I., Melancon G., Marshall S.: Graph Visualization and Navigation in Information Visualization: A Survey, Visualization and Computer Graphics, IEEE Transactions, vol. 6, Issue: 1, pp. 24-43 (2000).
9. Hobbs, M. H. W., Rodgers P. J.: Representing Space: A Hybrid Genetic Algorithm for Aesthetic Graph Layout, Frontiers in Evolutionary Algorithms, FEA'98, Proceedings of 4th Joint Conference on Information Sciences, JCIS'98, vol. 2, pp. 415-418 (1998).
10. Larrañaga P., Lozano J.A.: Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation, Kluwer Academic Publishers, 382 pages (2002).
11. Mühlenbein H., Mahnig T., Ochoa A.: Schemata Distributions and Graphical Models in Evolutionary Optimization, Journal of Heuristic, Vol. 5, No. 2, pp. 215-247 (1998).
12. Pach J., Tóth G.: Which crossing number is it, anyway? Proceedings of the 39th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Washington, DC, USA, pp. 617-626 (1998).
13. Rosete A., Ochoa A.: Genetic Graph Drawing, Proceedings of 13th International Conference of Applications of Artificial Intelligence in Engineering, AIENG'98, Adey, R. A., Rzevski, G., Nolan, P. (Ed.) Galway, Computational Mechanics Publications, pp. 37-40 (1998).
14. Rosete A.: Un Enfoque General y Flexible para el Trazado de Grafos, Tesis presentada en opción al grado científico de doctor en Ciencias Técnica, Facultad de Ingeniería Industrial, CEIS, La Habana, Cuba (2000).
15. Russell S., Norving P.: Artificial Intelligence : A Modern Approach, A Simon & Schuster Company Englewood Cliffs, New Jersey, Prentice Hall, pp. 111-114 (1995).

A New Approach to Music Information Retrieval using Dynamic Neuronal Networks

L.E. Gomez¹, J.H. Sossa¹, R. Barron¹, J.F. Jimenez¹,

¹ Centro de Investigación en Computación-IPN, Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n and M. Othón de Mendizábal, Zacatenco, México, DF. 07738, Mexico

sgomez08@sagitario.cic.ipn.mx, hsossa@cic.ipn.mx, rbarron@cic.ipn.mx, jfvielma@cic.ipn.mx

Abstract. The majority of work in music information retrieval (MIR) has been focused on symbolic representations of music. However, most of the digitally available music is in the form of raw audio signals. Although various attempts at monophonic and polyphonic transcription have been made, none has been successful and general enough to work with real world signals. So far, many researchers have been done to develop efficient music retrieval systems. In this paper, we develop a novel music retrieval system based on dynamic neural networks, which are trained with the signal melody, and not with traditional descriptors.

Keywords: Music Information Retrieval; Dynamic Neuronal Networks; musical descriptors.

1 Introduction

With the explosive expansion of digital music and audio contents, efficient retrieval of such data is getting more and more attention, especially in large-scale multimedia database applications. In the past, music information retrieval was based on textual metadata such as title, composer, singer or lyric. However, these various metadata-based schemes for music retrieval have suffered from many problems including extensive human labor, incomplete knowledge and personal bias.

Compared with traditional keyword-based music retrieval, content-based music retrieval provides more flexibility and expressiveness. Content-based music retrieval is usually based on a set of extracted music features such as pitch, duration, and rhythm.

In some works, such as [1][2], only pitch contour is used to represent melody. Music melody is transformed to a stream of U, D, R, which stands for a note is higher than, lower than, or equal to the previous note, respectively. But it simplifies the melody so much that it cannot discriminate music very well, especially when the music database is large.

In order to represent the melody more accurately and discriminatively, new feature sets have been proposed. In [3], pitch interval and rhythm are considered as well as

pitch contour. In [4], relative interval slope is used in music information retrieval. And [5] introduces four basic segment types (A,B,C,D) to model music contour.

When rhythm and pitch interval is considered, more complex similarity measure and matching algorithm should be used. [5] uses two-dimensional augmented suffix tree to search the desired song, rather than approximate string matching algorithm used in [1][2]. In [6], a new distance metrics between query and songs is proposed. But its computation is very time-consuming because it need adjust many parameters step by step to find the minimum distance.

Neural networks are characterized by dynamic dependence of events in past moments. Within the neural networks are dynamic networks are inherently dynamic, such as networks Hopfield, Jordan and Elman [1]. On the other hand, there are networks multilayer, which are static in nature but, achieve a dynamic behavior reinforced their own inputs samples of their previous outings.

In this paper, we propose a novel music retrieval system based on the use of dynamic neural networks, training with these melodies and using their synaptic weights as descriptors for the recovery of the melody.

The rest of this paper is organized as follows. In Section 2, we present an overview of ongoing research for analyzing music features and constructing MIR systems. In Sections 3, we describe our music retrieval system using dynamic neural networks. In Section 4, we report on some of the experimental results. Section 5 concludes this paper and describes our future directions.

2 Related work

In this section, we review some of typical techniques and systems for music information retrieval. As we know, music can be represented in two different ways. One is based on musical scores such as MIDI and Humdrum [7]. The other is based on acoustic signals which are sampled at a certain frequency and compressed to save space. Wave (.wav) and MPEG Layer-3 (.mp3) are examples of this representation.

2.1 Symbolic analysis

Many research efforts to solve the music similarity problem have used symbolic representation such as MIDI, musical scores, note lists and so on. Based on this, pitch tracking finds a "melody contour" for a piece of music. Next, a string matching technique can be used to compare the transcriptions of songs [1],[8],[9],[10],[11].

String matching has been widely used in music retrieval because melodies are represented using a string sequence of notes. To consider human input errors, dynamic programming can be applied to the string matching; however, this method tends to be rather slow. An inexact model matching approach [12] was proposed based on a quantified inexact signature-matching theory to find an approximate model to users' query requirements. It can enhance the reusability of a model repository and make it possible to use and manage a model repository conveniently and flexibly. Zhuge tried to apply this theory to a problem-oriented model repository system PROMBS [13].

There are also researches for symbolic MIR based on the ideas from traditional text IR. Using traditional IR techniques such as probabilistic modeling is described in [14] and using approximate string matching in [15]. Some work addressed other IR issues such as ranking and relevance. Hoashi [16] used relevance feedback for music retrieval based on the tree-structured vector quantization method (TreeQ) developed by Foote. The TreeQ method trains a vector quantizer instead of modeling the sound data directly.

2.2 Acoustic signal analysis

There are many techniques to extract pitch contour, pitch interval, and duration from a voice humming query. In general, methods for detecting pitches can be divided roughly into two categories: time-domain based and frequency-domain based.

In the time-domain, ZCR (zero crossing rate) and ACF (auto correlation function) are two popular methods. The basic idea is that ZCR gives information about the spectral content waveform cross zero per unit time [17]. In recent works, ZCR appeared in a different form such as VZCR (variance of ZCR) or SZCR (smoothing ZCR) [18]. On the contrary, ACF is based on the cross correlation function. While a cross correlation function measures the similarity between two waveforms along the time interval, ACF can compare one waveform with itself.

In the frequency-domain, FFT (fast Fourier transformation) is one of the most popular methods. This method is based on the property that every waveform can be divided into simple sine waves. But, a low spectrum rate for longer window may increase the frequency resolution while decreasing the time resolution. Another problem is that the frequency bins of the standard FFT are linearly spaced, while musical pitches are better mapped on a logarithmic scale. So, Forberg [19] used an alternative frequency transformation such as constant Q transform spectrums which are computed from tracked parts.

In recent works for the automatic transcription, they used probabilistic machine learning techniques such as HMM (hidden Markov model) and NN (neural network) to identify salient audio features and reduce the dimensionality of feature space. Ryyanen and Klapuri [20] proposed a singing transcription system based on the

HMM-based notes event modeling. The system performed note segmentation and labeling and also applied multiple-F0 estimation method [21] for calculating the fundamental frequency.

2.3 Recent MIR systems

For decades, many researchers have developed content based MIR (Music Information Retrieval) systems based on both acoustic and symbolic representations [1],[8],[22],[11].

Ghias [1] developed a QBH system that is capable of processing acoustic input in order to extract appropriate query information. However, this system used only three types of contour information to represent melodies. The MELDEX system [8] was designed to retrieve melodies from a database using a microphone. It first transformed

acoustic query melodies into music notations, and then searched the database for tunes containing the hummed (or similar) pattern. This web-based system provided several match modes including approximate matching for interval, contour, and rhythm.

MelodyHound [22], originally known as the "TuneServer", also used only three types of contour information to represent melodies. They recognized the tune based on error-resistant encoding. Also, they used the direction of the melody only, ignoring the interval size or rhythm. The C-BRAHMS [23] project developed nine different algorithms known as P1, P2, P3, MonoPoly, IntervalMatching, PolyCheck, Splitting, ShiftOrAnd, and LCTS for dealing with polyphonic music.

Suzuki [24] proposed a MIR system that uses both lyrics and melody information in the singing voice. They used a finite state automaton (FSA) as a lyric recognizer to check the grammar and developed an algorithm for verifying a hypothesis output by a lyric recognizer. Melody information is extracted from an input song using several pieces information of hypothesis such as song names, recognized text, recognition score, and time alignment information.

Many other researchers have studied quality of service (QoS)-guaranteed multimedia systems over unpredictable delay networks by monitoring network conditions such as available bandwidth. McCann [25] developed an audio delivery system called Kendra that used adaptability with a distributed caching mechanism to improve data availability and delivery performance over the Internet. Huang [26] presented the PARK approach for multimedia presentations over a best-effort network in order to achieve reliable transmission of continuous media such as audio or video.

3 Dynamic Neuronal Network applied to MIR

Dynamic neural network is the extension of static neural network via the consideration of time. The proposed dynamic models are developed based on static MLFN. In general, dynamics can be expressed by using a tapped-delay line, external dynamics and internal dynamics [27]. Tapped-delay line approach uses a sequence of delay to express dynamics and forms time-delay neural network [28],[29]. External dynamics approach uses the historical information of output itself to show dynamics and forms autoregressive type neural network [30],[31].

3.1 Multi-layer Feed-forward Neural Network (MLFN)

The most common network structure the MLFN which is a parallel distributed processing network. Parallel distributed processing network is formed by many basic units called neurons. Information processing takes place through the interactions of a large number of neurons can solve difficult tasks.

This is the idea of learning tasks via different angles via neurons and the interactions between neurons. It is a good option for model selection. Figure 1 shows the scheme network of MLFN.

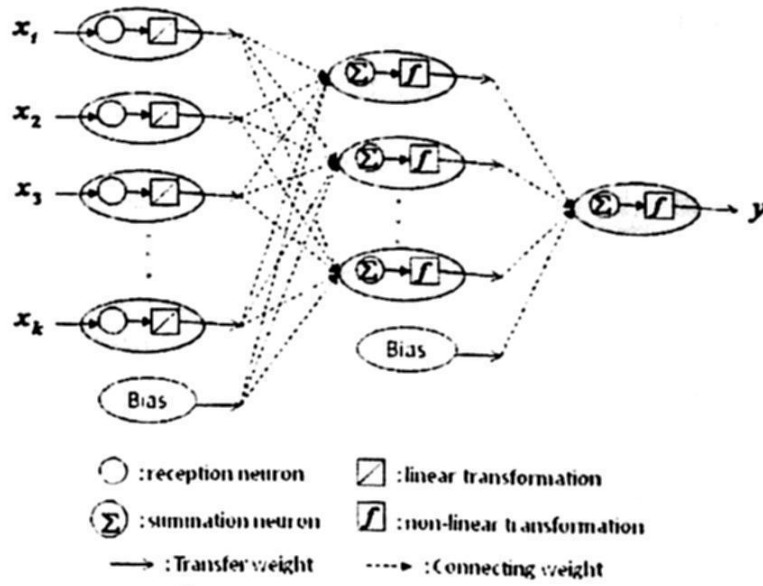


Fig. 1. MLFN Network Structure.

3.2 Time Delay Neural Network

Time delay neural network (TDNN) comes under dynamic neural networks, which are designed to explicitly include time relationships in the input-output mappings. Time-lagged feedforward networks (TLFNs) are a special type of dynamic networks that integrate linear filter structures inside a feedforward neural network to extend the non-linear mapping capabilities of the network with a representation of time [32]. Thus, in TLFN the time representation is brought inside the learning machine. The advantage of this technique is that the learning machine can use filtering information while the disadvantage is that the learning becomes complex since the time information is also coded in. TDNN is one of the specific cases of TLFN where a tapped delay line is given in the input followed by a multilayer perceptron (MLP) as shown in the block diagram in Fig. 2. Current input (at time t) and D delayed inputs (at time $t-1, t-2, \dots, t-D$) can be seen by the TDNN. The TDNN can be trained by using gradient descent back propagation. The ordered training patterns must be provided during training process [33].

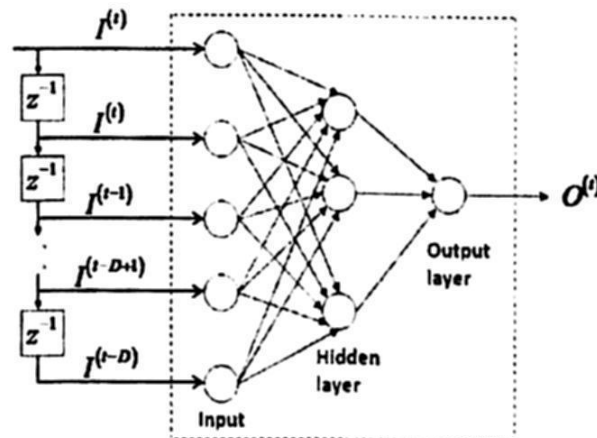


Fig. 2. Structure of Time Delay Neural Network.

3.3 Proposed method

It used WAV files, each file is trained in a dynamic neural network (TDNN), it shows a diagram in Figure 3. At the end of the training is obtained the weight matrix (WNN), as this is used as a descriptor of melody trained.

This method is novel because it works on the time domain, not you the frequency domain which gives a digital signature, such as: 1) Music features: pitch, duration, and rhythm. 2) Traditional descriptors: pitch contour, zero crossing rate, cross correlation, FFT, and others.

It is not necessary obtain the digital signature or features of the melody, because it is used in full. This reduces the level of a-priori knowledge of the melody by the user.

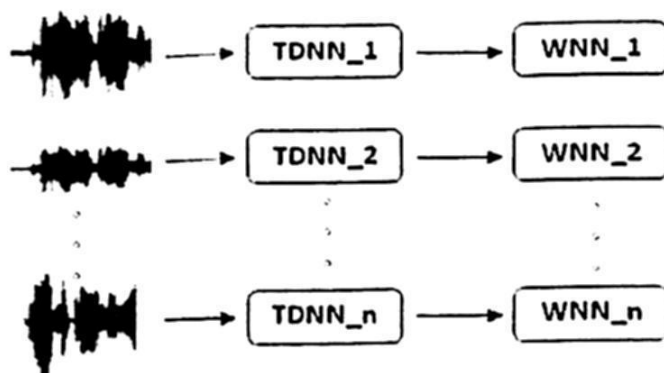


Fig. 3. Structure training of the TDNN with melodies.

The recovery of melodies is performed query with a segment of a melody, this segment is processed in the TDNN and stepdaughter and the descriptors of the melodies, get the error recovery the melody, and finally with the argument minimum, you get the index gives melody that was query, it shows a diagram in Figure 4.

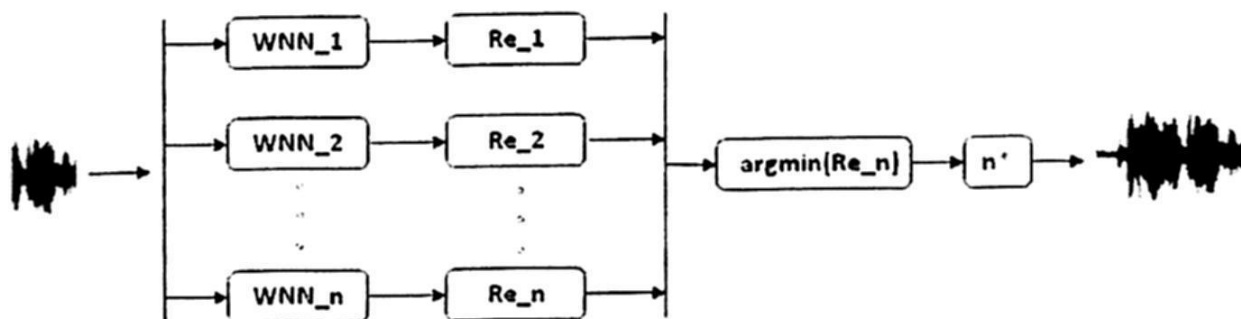


Fig. 4. Structure retrieval of melodies using descriptors of the TDNN.

4 Experimental results

It used 16-bit WAV files (stereo mode), each file is trained in a dynamic neural network, and these networks have a maximum of 100 iterations, and 10 neurons in the hidden layer. At the end of the training is obtained the weight matrix, as this is used as a descriptor of melody trained.

Tests were with different numbers of neurons in the hidden layer, as well as different numbers of iterations. the results of this test are shown in Tables 1.2 and Figures 5.6.7.8. The compares the errors rate training and recovery.

Table 1. Table of error rate training and recovery, with different numbers of neurons

N. neurons	Training			Recovery		
	Minimum	Average	Maximum	Minimum	Average	Maximum
5	2.99E-04	2.48E-03	6.12E-03	8.45E-03	1.41E-02	2.21E-02
6	2.93E-04	2.48E-03	5.61E-03	5.19E-03	1.88E-02	5.05E-02
7	5.59E-04	3.67E-03	6.46E-03	6.09E-03	1.95E-02	4.83E-02
8	2.94E-04	3.52E-03	6.43E-03	1.01E-02	1.69E-02	3.00E-02

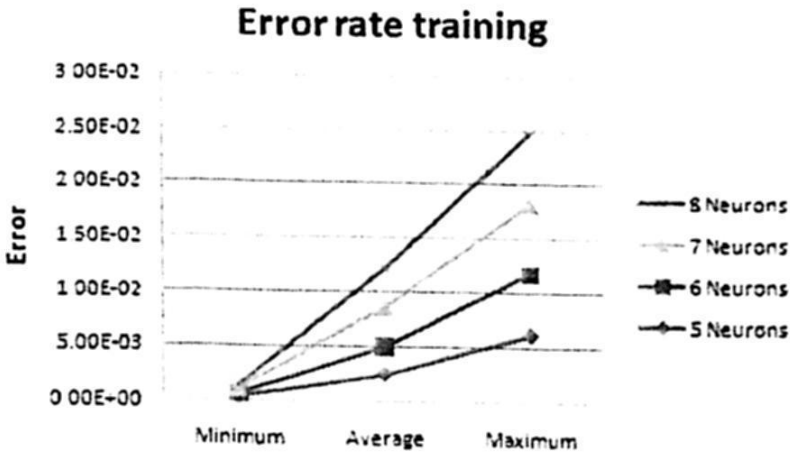


Fig. 5. Graphic of error rate training, with different number of neurons.

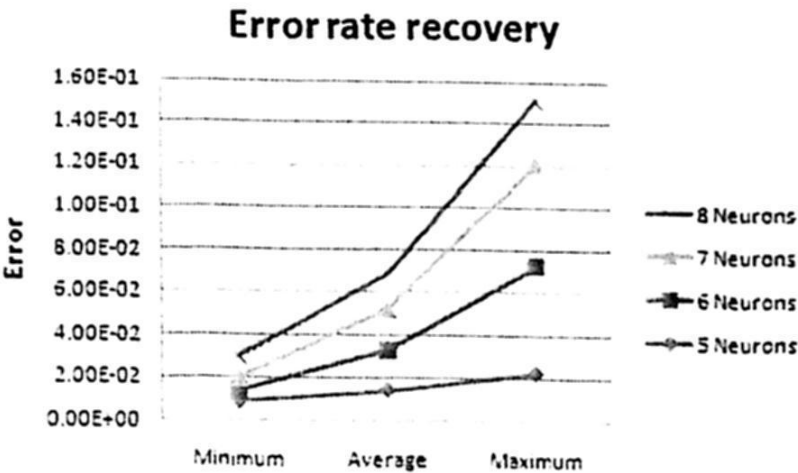


Fig. 6. Graphic of error rate recovery, with different number of neurons.

Table 2. Table of error rate training and recovery, with different numbers of iterations

N. iterations	Training			Recovery		
	Minimum	Average	Maximum	Minimum	Average	Maximum
10	2.07E-03	9.67E-03	2.64E-02	9.96E-03	3.00E-02	6.37E-02
25	1.15E-03	5.65E-03	1.09E-02	5.49E-03	2.14E-02	5.68E-02
50	2.99E-04	2.48E-03	6.12E-03	8.45E-03	1.41E-02	2.21E-02
75	3.55E-04	5.29E-03	1.05E-02	7.94E-03	2.11E-02	5.08E-02

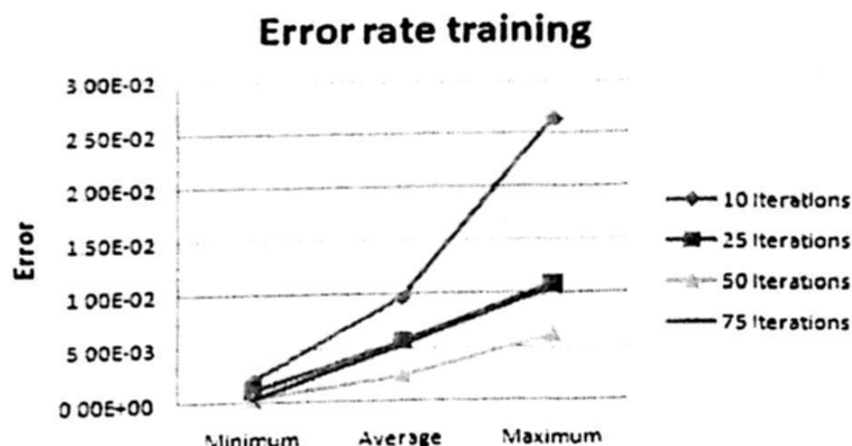


Fig. 7. Graphic of error rate training, with different number of iterations.

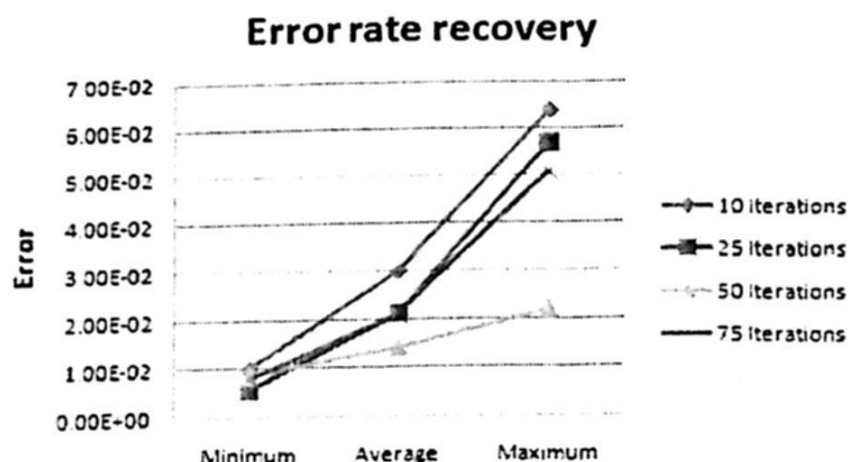


Fig. 8. Graphic of error rate recovery, with different number of iterations.

The system works effectively and efficiently, with a query with a segment less than 1 percent the total melody, the melody is recovered.

For example, if you train the TDNN with a melody the 132.221 frames, and recovers with only 350 frames, which is 0.264 percent of the total of the melody, with these we can conclude that is recovering with less than 1 percent of the melody.

4 Conclusions

Content-based music retrieval is a very promising method for large music library, yet it is a very challenging task. We discussed various features of music contents for content-based retrieval.

In this paper, we have presented a preliminary approach to Music Information Retrieval. The goal of this study was to explore a new line research within the field of MIR. Not all people are experts in models auditory perception, so we have chosen a TDNN network type that is capable of solving the problem from the samples without using any traditional descriptor or digital signature. Neither has made any preprocessing before the music files, these apply changes a melody can distort or

highlight the information contained therein. Therefore, unlike other techniques, MIR, we have original melody introduced directly into the net try find this as a suitable height relations present in the spectrum signal.

The inputs to TDNN network as a series of amplitudes obtained from stereo melody. The output of network is the encoding of a musical descriptor in a matrix of weights.

The system works effectively and efficiently, as a query with a segment less than 1 percent the total melody, the melody is recovered.

It can be concluded that the system retrieval using dynamic neural network is a success, achieving very faithful to the melody identification, in any case, the results of this study open many lines promising for further research on MIR by dynamic neural networks.

Acknowledgements. We wish to thank the Centro de Investigación en Computación of the I.P.N. by the support to accomplish this project. L.E. Gomez and J.F. Jimenez thanks CONACYT by the scholarship received to complete his doctoral studies. R. Barron thanks the SIP-IPN under grant 20100379 for the support. H. Sossa thanks the SIP-IPN under grant 20091421 for the support. H. Sossa also thanks CINEVESTAV-GDL for the support to do a sabbatical stay from December 1, 2009 to May 31, 2010. Authors thank the European Union, the European Commission and CONACYT for the economical support. This paper has been prepared by economical support of the European Commission under grant FONCICYT 93829. The content of this paper is an exclusive responsibility of the CIC-IPN and it cannot be considered that it reflects the position of the European Union. Finally, authors thank the reviewers for their comments for the improvement of this paper.

References

- [1] Ghias, A.: Query By Humming-Musical Information Retrieval in an Audio Database. Proc.s of ACM Multimedia 95, pp231-236, 1995.
- [2] Blackburn, S., De Roure, D.: A Tool for Content Based Navigation of Music. Proc. ACM Multimedia 98, pp 361-368, 1998.
- [3] McNab, R.J.: Towards the Digital Music Library: Tune Retrieval from Acoustic Input. Proc. of Digital Libraries, pp 11-18, 1996.
- [4] Lemstrom, K., Laine, P., Perttu, S.: Using Relative Interval Slope in Music Information Retrieval. In Proc. of International Computer Music Conference 1999 (ICMC '99), pp. 317-320, 1999.
- [5] Chen, A.L.P., Chang, M., Chen, J.: Query by Music Segments: An Efficient Approach for Song Retrieval. In Proc. of IEEE International Conference on Multimedia and Expo., 2000.
- [6] Francu, C. Nevill-Manning, C.G.: Distance Metrics and Indexing Strategies for a Digital Library of Popular Music. In Proc. of IEEE International Conference on Multimedia and Expo. 2000.
- [7] Kornstadt, A.: Themefinder: A web-based melodic search tool. In: Computing in Musicology 11. MIT Press., 1998.
- [8] McNab, R.J. et al.: The New Zealand digital library melody index. Digital Libraries Magazine., 1997.

- [9] Uitdenbogerd, A., Zobel, J.: Melodic matching techniques for large music databases. In: *Proceedings of ACM Multimedia Conference*, pp. 57–66., 1999.
- [10] Hwang, E., Rho, S.: FMF(fast melody finder): A web-based music retrieval system. In: *Lecture Notes in Computer Science*, vol. 2771. Springer-Verlag, pp. 179–192., 2004.
- [11] Hwang, E., Rho, S.: FMF: Query adaptive melody retrieval system. *Journal of Systems and Software* 79 (1), 43–56. 2006.
- [12] Zhuge, H.: An inexact model matching approach and its applications. *Journal of Systems and Software* 67 (3), 201–212. 2003.
- [13] Zhuge, H.: A problem-oriented and rule-based component repository. *Journal of Systems and Software* 50 (3), 201–208. 2000.
- [14] Pickens, J.: A comparison of language modeling and probabilistic text information retrieval approaches to monophonic music retrieval. *Proceedings of the 1st Annual International Symposium on Music Information Retrieval (ISMIR2000)*, 2000.
- [15] Lemstrom, K., Wiggins, G.A., Meredith, D.: A threelayer approach for music retrieval in large databases. In: *Second International Symposium on Music Information Retrieval*. Bloomington, IN, USA, pp. 13–14. 2001.
- [16] Hoashi, Matsumoto, Inoue.: Personalization of user profiles for content-based music retrieval based on relevance feedback. *ACM Multimedia*, pp. 110–119. 2003.
- [17] Gerhard, David.: Pitch Extraction and Fundamental Frequency: History and Current Techniques. Technical Report TR-CS 2003-06. 2003.
- [18] Huang, R., Hansen, J.H.L.: Advanced in unsupervised audio classification and segmentation for the broadcast news and NGSW Corpora. *IEEE Trans. on Audio, Speech and Language Processing* 14 (3), 907–919. 2006.
- [19] Forberg, Johan.: Automatic conversion of sound to the MIDIformat. TMH-QPSR 1-2/1998. 1998.
- [20] Ryynanen, Matti., Klapuri, Anssi.: Transcription of the singing melody in polyphonic music. *ISMIR 2006*. 2006.
- [21] Klapuri, Anssi P.: A perceptually motivated multiple-f0 estimation method. 2005 IEEE workshop on applications of signal processing to audio and acoustics, 291–294. 2005.
- [22] Typke, R., Prechelt, L.: An interface for melody input. *ACM Transactions on Computer-Human Interaction*., 133–149. 2001.
- [23] Ukkonen, E., Lemstrom, K., Makinen, V.: Sweepline the music. *Lecture Notes in Computer Science* 2598, 330–342. 2003.
- [24] Motoyuki Suzuki, et al.: Music information retrieval from a singing voice based on verification of recognized hypothesis. *ISMIR 2006*. 2006.
- [25] McCann, J.A. et al.: Kendra: Adaptive Internet system. *Journal of Systems and Software* 55 (1), 3–17. 2000.
- [26] Huang, C.M. et al.: Synchronization and flow adaptation schemes for reliable multiple-stream transmission in multimedia presentation. *Journal of Systems and Software* 56 (2), 133–151. 2001.
- [27] Nelles, O.: *Nonlinear system identification*. Springer, Germany. 2001.
- [28] Yun, S.Y., Namkoong, S., Rho, J.H., Shin, S.W. and Choi, J.U.: A performance evaluation of neural network models in traffic volume forecasting, *Mathematic Computing Modelling*, Vol. 27, No.9-11, 293-310. 1998.
- [29] Lingras, P. and Mountford, P.: Time delay neural networks designed using genetic algorithms for short term inter-city traffic forecasting. In L. Monostori, J. Vancza and A. Moonis (eds.), *IEA/AIE 2001*. Springer, Berlin. 2001.
- [30] Campolucci, P., Uncini, A., Piazza, F. and Rao, B.D.: On-line learning algorithms for locally recurrent neural networks, *IEEE transactions on neural networks*, Vol. 10, No. 2, 253-271. 1999.

- [31] Tsai, T-H., Lee, C-K. and Wei, C-H.: Artificial Neural Networks Based Approach for Short-term Railway Passenger Demand Forecasting, *Journal of the Eastern Asia Society for Transportation Studies*, Vol. 4, 221-235, 2003.
- [32] Haykin, S.: *Neural Networks: A Comprehensive Foundation*, 2nd Edition, Prentice Hall PTR. ISBN 0-13-273350-1, p. 837, 1998.
- [33] Weibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.: Phenomena Recognition Using Time-delay Neural Networks. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37, pp. 328-339, 1989.

Conformal geometric algebra and sphere fitting applied to colour image segmentation

Luis Horna, Ricardo Barrón and Giovanni Guzmán

Artificial Intelligence Laboratory
Centro de Investigación en Computación
Mexico D.F, Mexico

chornab08@sagitario.cic.ipn.mx, {rbarron, jguzman1}@cic.ipn.mx

Abstract. A common task in computer vision is to segment regions based in its colour. Over the years several algorithms have been proposed, ranging from thresholding to more sophisticated algorithms such as belief propagation. In this paper, conformal geometric algebra is used to fit spheres to a set of points representing colours of some colour space, showing how these spheres can be used to identify pixels in images that belong to the covering spheres, i.e segment regions based in its colour.

1 Introduction

In general, the conformal model can be seen as an extension of the Euclidean space R^n , where the conformal model is generated by $\{e_o, e_\infty, e_1, \dots, e_n\}$, where $e_o \cdot e_i = e_\infty \cdot e_i = 0$, $e_o^2 = e_\infty^2 = 0$ and $e_o \cdot e_\infty = -1$. In the conformal domain a Euclidean point x is represented as:

$$X = x + \frac{1}{2}x^2e_\infty + e_o \quad (1)$$

and spheres are represented by:

$$S = C - \frac{1}{2}\gamma^2e_\infty \quad (2)$$

where C is a conformal point that represents the centre of S and γ is the radius.

In the conformal model distance from a point to a sphere is given by:

$$X \cdot S = \gamma^2 - (x - c)^2 \quad (3)$$

Eq.(3) is equivalent to the concept of "power" between a point and sphere, which is used in the context of covering spheres to know how much a point belongs to a sphere.

From this point of view, the segmentation of images based on its colour can

be expressed as fitting a sphere S to a set X of colours that are similar, for instance a certain tone of red.

Stating the problem in a geometric way makes possible to solve it by just computing the "power" between the S and some point Y , fig.(1) illustrates three different situations of how spheres can be used to segment colour images.

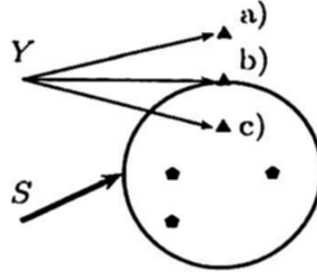


Fig. 1. Three possible situations used to segment regions based on its colour

where:

- a) if $Y \cdot S < 0$ then Y is outside S .
- b) if $Y \cdot S = 0$ then Y lies on S .
- c) if $Y \cdot S > 0$ then Y is inside S .

one clear restriction is that colours contained by S must be similar, because using colours which are very different would result in a sphere that would not be useful. In such situation it is preferable to use several spheres containing different colours.

2 Fitting spheres to colour points

In recent years, treating the colour pixels as points in R^3 regardless of their colour space has been used for gradient detection [1], [2], segmentation [3], and colour representation [5]. Representing colour pixels in such a way makes the problem of segmentation suitable to be stated as a geometric problem, where Geometric Algebra (GA) becomes a powerful tool.

Fitting a sphere S to some $X \subset R^n$ consists in finding S such that it optimally covers all points in X . From this statement, spheres become an interesting option to represent colours which are "similar".

This concept can also be used to fit a sphere to a set of colour pixels $X = \{X_1, \dots, X_m\} \in R^3$ using a least square approach:

$$S_* = \underset{S}{\operatorname{argmin}} \sum_{i=1}^m (X_i \cdot S)^2 \quad (4)$$

where $X_i \cdot S$ can be represented in a matrix form as follows

$$W = \begin{bmatrix} X_1^1 & X_1^2 & X_1^3 & -1 & \frac{1}{2}X_1 \cdot X_1 \\ X_2^1 & X_2^2 & X_2^3 & -1 & \frac{1}{2}X_2 \cdot X_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_n^1 & X_n^2 & X_n^3 & -1 & \frac{1}{2}X_n \cdot X_n \end{bmatrix} \quad (5)$$

where X_i^j represents the j -th component of the i -th point in $X \in R^3$, eq.(5) can be solved by computing the eigen-values, and eigen-vectors [6] of $B = W'W$.

Once S has been computed it can be used to determine whether some point Y is found inside S , this is done in the following way:

- if $Y \cdot S \geq 0$ then Y is inside S .
- if $Y \cdot S < 0$ then Y is outside S .

an advantage of computing S is that it can be used on different images to segment colours. It should be noted that sometimes a point Y that is similar to those in X could be outside S , in which case it is preferable to use a threshold t .

3 Experimental results

In this section we show to different experiments of how the proposed solution works. First we use the well known image of peppers to segment regions of similar colour, then using real image of human retina, the optic disk is segmented. The second experiment consists in segmenting objects of different colours.

Fig.(2(a)), shows the image before segmentation, in order to create a sphere S a set of colours X must be created, this is done manually by selecting some point from the area of interest, in this experiment only four samples were taken from the red pepper (right side in fig.2(a)), the result of the segmentation process using the different thresholds are shown in fig.(3).

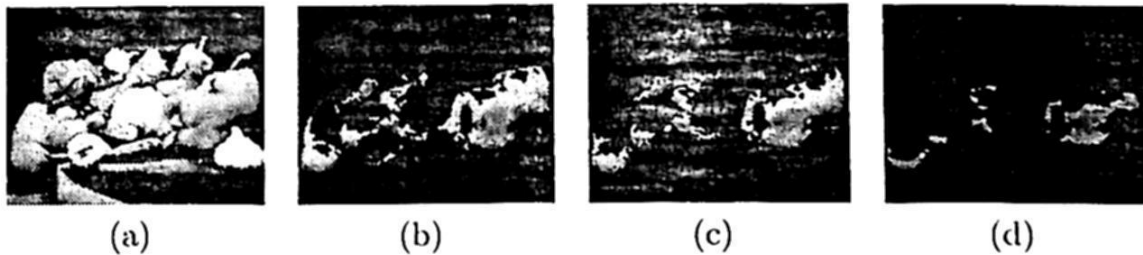


Fig. 2. Result of using different thresholds -0.1 , -0.05 and 0.0 for images (b) , (c) and (d) respectively

the proposed solution was also used to segment the optic disk from human retina images, the images were also sampled over the optic disk, fig.(3), only five samples were taken to create a sphere before segmenting the optic disk.

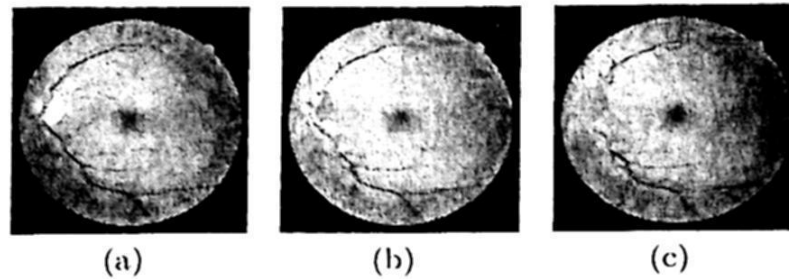


Fig. 3. (a) original image, (b) segmentation using a threshold of -0.1 and (c) segmentation using a threshold of -0.2

The second experiment consisted in segmenting objects of different colours, unlike the previous experiment where only one sphere was used, in this experiment several spheres corresponding to various colours are used, fig.(3), five samples per colour are taken before segmenting a region of the selected colours, then separate spheres are created for each selected colour.

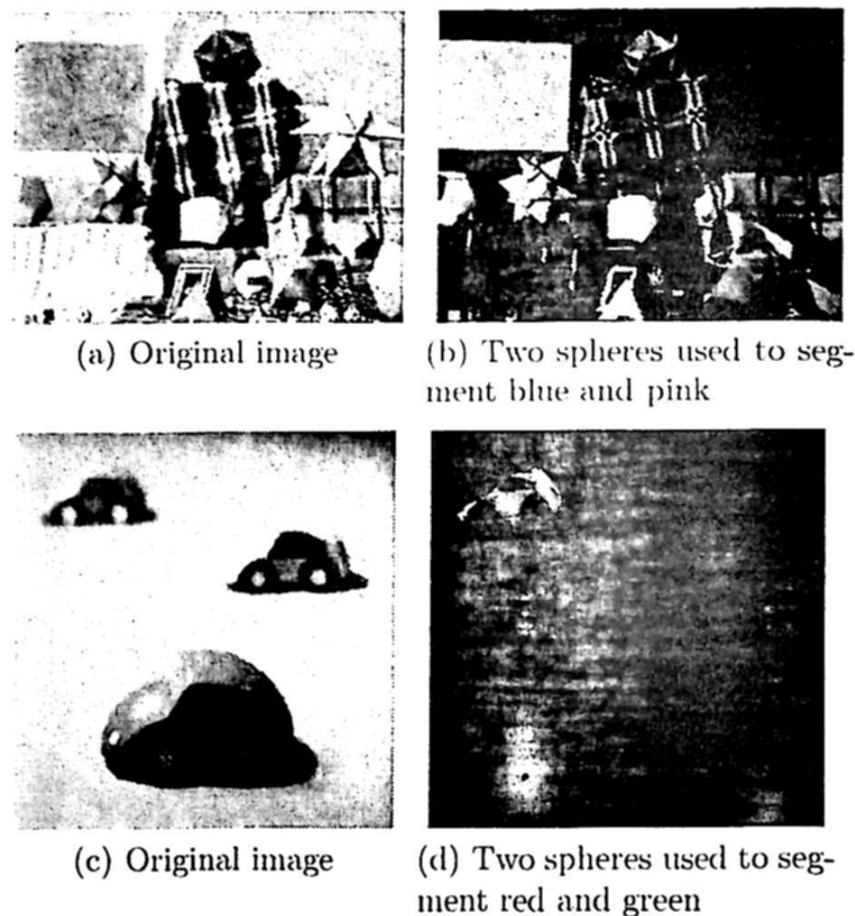


Fig. 4. Result of different spheres for colour segmentation

This last experiment had some problems segmenting different regions simultaneously, because colours were relatively close. This situation favoured negative thresholds, which lead to situations such as in fig.(4(d)), where yellow colour was segmented when only red and green were intended to be segmented.

In these two experiments we have chosen to take only four or five samples of a given colour, it is not a restriction of the proposed method. However, it should be noted that it is preferable to select colours that are very similar.

4 Future work

The method presented in this paper has the limitation of using spheres that contain colours that must be similar, which sometimes results in difficulties to establish a reliable threshold. Therefore one task that will be addressed in the future is to determine a maximum and minimum threshold, which should take into account the properties of the spheres that represent some colour. It would also be interesting to adapt our proposed solution to other segmenting tasks such as stereo matching in multiple views and motion tracking.

5 Conclusion

In this paper it has been shown how the colour segmentation problem can be stated in geometric way, and the useful tool GA can be in solving problems of computer vision. It must also be noted that using the proposed solution to segment real images is a good option since it allows to take samples from different images and use the computed sphere on images that were not used to take samples.

References

1. A. Cumani (1991), Edge detection in multi-spectral images. *Comput. Vis. Graph. Image Process. Graph. Models Image Process.* 5, pp 4051.
2. Thomas Batard, Christophe Saint Jean and Michel Berthier (2008). A metric Approach to nD images edge detection with Clifford algebras, In: *J. Math Imaging Vis.*, pp 296-312, Norwell USA: Kluwer Academic Publishers.
3. G. Urcid, J.C Valdiviezo-N and G. Ritter (2009), Color Image Segmentation based on lattice algebra auto-associative memories, In: *Proceedings of ASC 2009*, Calgary: ACTA Press.
4. Christian Perwass (2009). *Geometric Algebra with Applications in Engineering*, London: Springer Verlag.
5. Jesús Angulo (2010). Geometric algebra colour image representations and derived total orderings for morphological operators - Part I: Colour quaternions, *J. Vis. Comun. Image Represent.*, pp 33 - 48, Orlando USA: Academic Press.
6. Dietmar Hildenbrand, *Geometric Computing in Computer Graphics using Conformal Geometric Algebra*, Interactive Graphics Systems Group, TU Darmstadt, Germany.

Algorithm of support for the detection of the Acute Lymphoblastic Leukemia

Susana Ordaz Gutiérrez¹, Fabián Torres Robles¹,
Francisco Javier Gallegos Funes², Alberto Jorge Rosales Silva².

¹ Estudiante de Maestría en Electrónica, SEPI -ESIME-IPN, México D.F., México

² Profesor-Investigador, SEPI-ESIME-IPN, México D.F., México.
ESIME-SEPI-IPN, México D.F., Unidad Profesional Adolfo López Mateos,
Edif. 5, 3er Piso, Col. Lindavista, C.P. 07738, México, D.F.

Tel. +Fax: 52(55)5729-6000 ext. 54608

susana.ordaz@gmail.com

Abstract. This project bases on the worry that Acute Lymphoblastic Leukemia Infantil is the most common type of cancer in children, generally it deteriorates rapidly but it can be treated in time. ALL is a disease in which white blood cells attack the infections (so called lymphocytes), which are immature in big quantities in the blood and bony marrow of the child. There is an algorithm designed that helps detect the cancer ALL, this one must be a support for the doctor. The samples that are taken of bony marrow dress in the microscope. Our algorithm is based on the programming on Mat lab, this one consists of four stages of processing: Image of entry, Segmentation, Classification, Recognitions and Exit. First the image that is selected of a bank of information is read, where all our images of tests are read as well, the original image turns into a scale of gray image, in the stage of segmentation 6 Edge's methods are analyzed to see which is most indicated for our needs, an object can be discovered easily in an image if the object has the contrast of sufficient bottom, this is done in the same stage of segmentation.

Keywords: Acute Lymphoblastic Leukemia (ALL).

1 Introduction

ALL is a type of cancer for which the bony marrow produces too many lymphocytes (a type of white blood cell).

The ALL is a cancer of the blood and the bony marrow. This type of cancer generally deteriorates rapidly if it is not treated fast. It is the most common type of cancer in the children. Normally, the bony marrow mother elaborates blood cells that turn, with the time, into blood mature cells.

The mother cell-amyloid turns into one of three types of blood mature cells:

1. Red blood cells that transport oxygen to all the fabrics of the body.
2. White blood cells that fight against the infections and the diseases.
3. Platelets that help to anticipate hemorrhages that form clots of blood.

In the ALL, too many mother cells can turn into a type of white blood cells called lymphocytes. These lymphocytes also call lymphoblast or leukemia cells. There are three types of lymphocytes:

Lymphocytes B that produce antibodies to help to fight against the infections.
Lymphocytes T that help lymphocytes B to generate the antibodies that help fight against the infections. Aggressive natural cells that attack the cancerous cells or the virus.

2 Development

Evolution of a blood cell.

A mother blood cell goes through several stages to turn into a red blood cell, a platelet or a white blood cell.

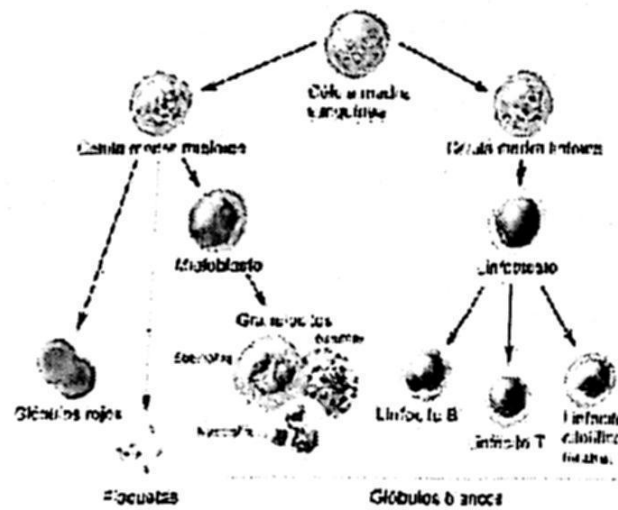


Fig. 1. Evolution of a blood cell.

In the ALL case, the lymphocytes cannot fight very well against the infections. In addition, as it increases the quantity of lymphocytes in the blood and the bony marrow, there is less place and space for the white blood cells, the red blood cells and the healthy platelets. This can lead to infections, anemia.

There are subgroups of child ALL.

Four of the subgroups of child ALL are based on the type of blood cells that are affected, if they present certain changes in the chromosomes and the age in the moment of the diagnosis:

ALL of cells T.

Positive ALL for the chromosome Philadelphia.

ALL diagnosed in a breast-fed baby.

ALL diagnosed in 10-year-old children of age or more, and in teenagers.

The exhibition to radiation and the family precedents can take part in the risk of suffering from child ALL.

A factor of risk is anything that increases the risk of contracting of a disease. People who have a factor of risk lets us know that one is going to contract cancer; not to have a factor of risk means that one is not going to contract cancer. The people who think that they can be in risk, must consult this topic with their doctor. The possible factors of risk for the ALL include the following aspects:

- To have a brother with leukemia.
- To be of white race or of Hispanic origin.
- To reside in the United States of America.
- To be exposed to the X-rays before the birth.
- To be exposed to radiation.
- To have had a previous treatment with chemotherapy or other medicines that debilitate Immunological system
- To suffer from certain genetic disorders as Down's syndrome.
- The possible signs of infantile ALL include fever and bruises.

These and other symptoms can be caused by the infantile ALL. Other affections can cause the same symptoms. It must consult with a doctor if one presents any of the following problems:

- Fever.
- Bruises or bled easy.
- Petequia (flat spots, like dots under the skin produced by the bled one).
- Aching bones or joints.
- Masses that do not hurt in the neck, the armpits, the stomach or the groin.
- Pain or sensation of satiety under the ribs.
- Weakness or sensation of weariness.
- Loss of appetite.
- To detect and diagnose child ALL, there are tests that examine the blood and the bony marrow.

You can use the following tests and procedures:

1. Physical examination and precedents: examination of the body to check the general signs of health, inclusive the checkup of signs of disease, as masses or any other thing that seems to be abnormal. There take also the medical precedents of the diseases and the previous treatments of the patient.

2. Blood complete inventory (RSC) with differential: procedure by means of which a sample of blood is taken and the following aspects are analyzed:

- The quantity of red blood cells and platelets.
- The quantity and the type of white blood cells.
- The quantity of hemoglobin in the red blood cells.
- The part of the sample composed by red blood cells.

3. Aspiration of bony marrow and biopsy: extraction of a sample of bony marrow, blood, and a small chunk of bone, this is done with the insertion of a needle in the bone of the hip or the breastbone. A pathologist observes the samples of bony marrow, blood and bone under a microscope to check if there are signs of cancer.

4. Analysis cytogenetic: this procedure is done by observing under a microscope the cells of the blood sample or the bony marrow to check if there are certain changes in the chromosomes of the lymphocytes. For example, in the ALL, part of a chromosome moves to another chromosome. This is called "chromosome Philadelphia"

5. Immunophenotyping: this procedure is done by observing under a microscope the cells of the blood sample or the bony marrow to check if the malignant (cancerous) lymphocytes started by being lymphocytes B or lymphocytes T.

6. Studies of the chemistry of the blood: procedure in which a sample of blood is examined to measure the quantities of certain substances liberated to the blood for the organs and fabrics of the body. A slightly common quantity of a substance can be a sign of disease in the organ or the fabric that elaborates it.

7. X-ray photography of thorax: X-ray photography of the organs and bones of the interior of the thorax. An X-ray is a type of bundle of energy that can cross the body and to take form of a movie that shows an image of the interior of the body.

Certain factors affect the forecast (possibility of recovery) and the options of treatment. The prediction and the options of treatment depend on the following aspects:

The age and inventory of white blood cells in the moment of the diagnosis.
How rapid and how much diminishes the concentration of leukemia cells after the initial treatment.

The kind and the ethnic race.

If the leukemia cells originated in lymphocytes B or in lymphocytes T.

If certain changes demonstrated in the chromosomes of the lymphocytes.

If the leukemia has been spread up to the brain and the spinal marrow.

If the child suffers from Down's syndrome.



Fig. 2.

a) Healthy blood cell

b) Blood cell with ALL

2 Stages of processing

Where do we want to get?

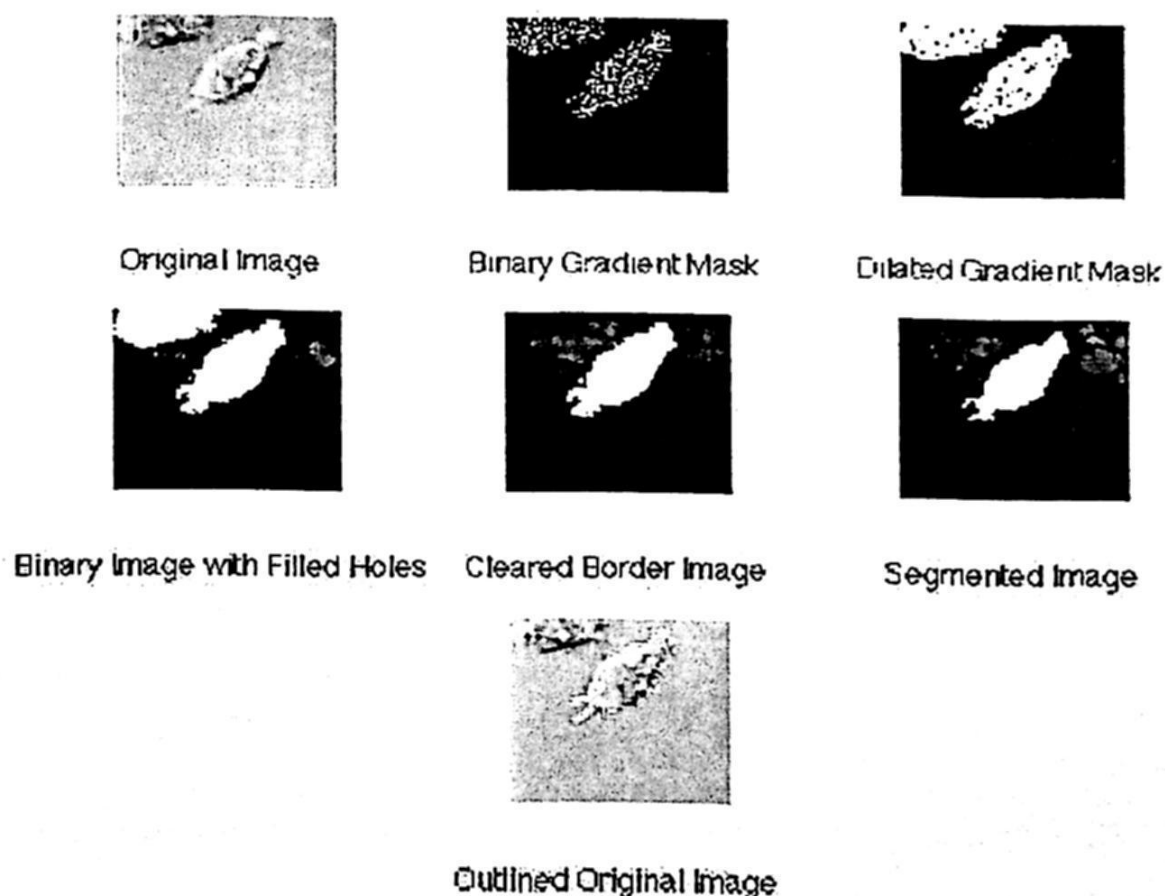


Fig. 3. To obtain a good segmentation in the images, there are methods of contour detection and basic instruments of morphology that later I will describe.

3 To read the image

In mat lab a scale of gray image is represented by a two-dimensional counterfoil of $m \times n$ elements where n represents the number of pixels of width and m the number of pixels of length. The element v_{11} corresponds to the element of the top left corner, where every element of the counterfoil of the image has a value of 0 (black) to 255 (white).

To read images contained in a file to the environment of mat lab the function is in use `imread`, whose syntax is `imread('name of the file')`.

Where name of the file is a chain of characters containing the complete name of the image with its respective extension, the formats of images that are supported by mat lab.

To introduce an image saved in a file with one of the formats specified in the previous table, only the function has to be used `imread` and assign its result to a variable that will represent the scale of gray image.

Formato	Extension
TIFF	.tiff
JPEG	.jpg
GIF	.gif
BMP	.bmp
PNG	.png
XWD	.xwd

Fig. 4. Format of image

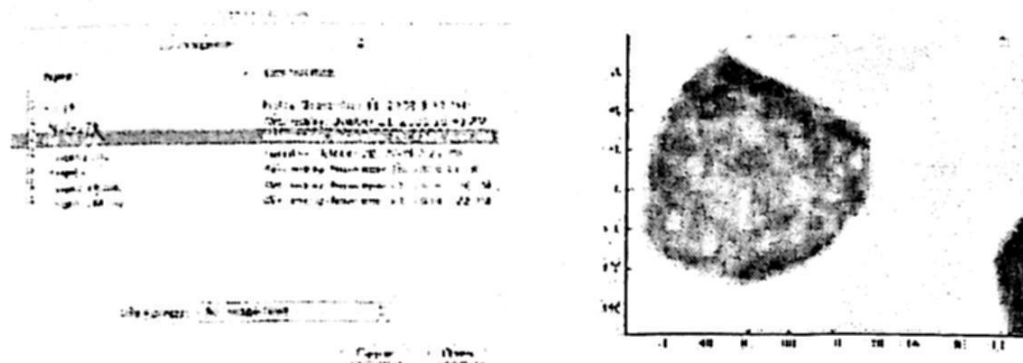


Fig. 5. An image is selected inside the bank of information that wants to be analyzed.

4 Segmentation (Detection of edges)

We spend to the second part of the stages of processing.

Functions for the extraction of edges

In a computer like vision we proceed to the recognition of objects or segment regions, to extract the edges of objects (that theoretically delimit its sizes and regions). The function `edge` gives the possibility of obtaining the edges of the image. The function allows to find the edges from two different algorithms that can be chosen, *canny* and *sobel*. The format of this function is:

$$ImageT=edge(ImageS, algorithm); \quad (1)$$

Where `ImageT` is the image obtained with the extracted edges, `ImageS` is the variable that contains the image in scale of gray to which one tries to recover its edges, whereas `algorithm` can be one of the two *canny* or *sobel*. In such a way that if

to the image in scale of gray contained in the variable image gray its edges they want to recover him using in algorithm canny that would be written in line of commands:

$$\text{ImageR}=\text{edge}(\text{imagegray}, \text{canny}); \quad (2)$$

Definition of edge: to the pixels where the intensity of the image changes abrupt form. Significant reduction of the quantity of information and it leaks or filters the unnecessary information. It is the calculation of a local operator of derivation.

We detect edges using the operators of derivation:

Using the first derivative we have:

- Positive: change the levels of gray so the image is more clear.
- Negative: opposite case.
- Zero in zone of uniform grey.
- And for the second derivative we have:
- Positive value: dark zone of every edge.
- Negative value: clear zone of every edge.
- Value zero: gray zone is a constant value.

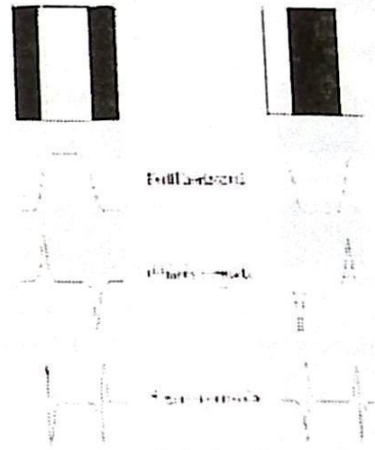


Fig. 6. We detect edges using the operators of derivation:

The value of the magnitude of the first derivative serves us to detect the presence of edges. The sign of the second derivative indicates us if the pixel belongs to the clear zone or to the dark zone. The first derivative in any point of the image will be given by the magnitude of the gradient the second derivative will be given by the operator Laplacian.

5 Edge's Methods.

EDGE finds edges in intensities of the image. EDGE takes an intensity or a binary image and returns a binary image BW of the same size that where the function finds edges. EDGE supports six different methods that find edge:

The Sobel method finds edges that use the approximation Sobel to the derivative. This returns edges in those points where the gradient of I is the maximum.

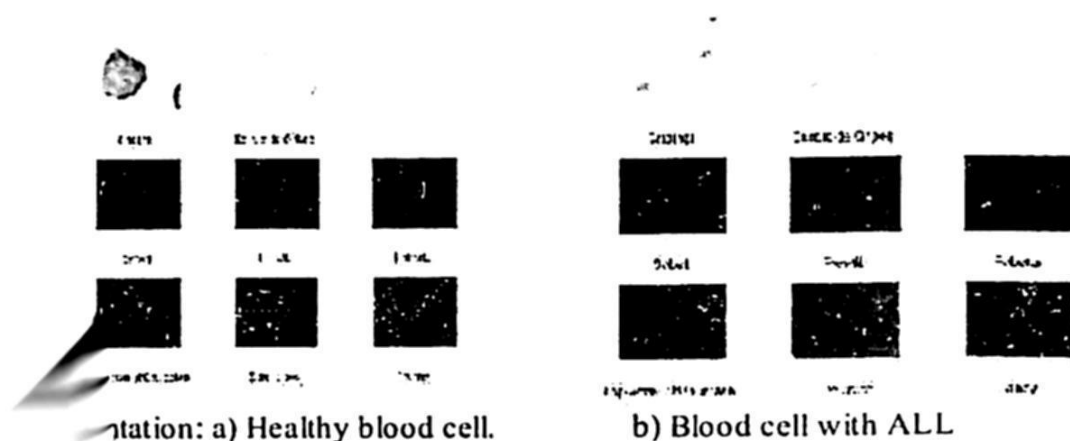
The Prewitt method finds edges that use the approximation Prewitt to the derivative. This returns edges in those points where the gradient of I is the maximum.

The Roberts method finds edges that use the approximation Roberts to the derivative. This returns edges in those points where the gradient of me is the maximum.

The Laplacian of Gaussian method finds edges for search cross zero after the filtration I with a Laplacian of Gaussian filter.

The method Zero-cross-country race finds edges looking for crossings for zero after the filtration I with a filter that you specify. The method Canny finds edges of the local maxim of the gradient of I . The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds, to discover strong and weak edges, and includes the weak edges in the exit only if they are connected to strong edges. This method is therefore less probable than the others to be "cheated" by the noise, and more probably to discover real weak edges.

The parameters that are possible to give differ according to the method that you specify. If there is no specified method EDGE, it uses the method Sobel.



6 am

The histogram is the representation of the density of probability of every value of grey for this image.

Both the histogram, and the histogram equalization, are vectors.

Equalizer the Histogram is to do everything possible to flat and separate everything. This is what the pixels do to distribute a whole range of values (from 0 to 255) and that in the image equalized will highlight details that before were not evident.

To generate an image with the histogram equalized several steps are needed:

1. The histogram of the image is calculated
2. To normalize the histogram (to divide it between the total number of pixels).
3. To calculate the histogram accumulated (the pixels to be added from the value 0 to 255, this will originate an increasing graph)
4. The application of the algorithm (since it is a question of counterfoils, it must be inside two sheltered curls):

For values of the original image different from zero:

$Imagen_ecualizada(i,j) = histograma_acumulado(imagen_original(i,j))$ (3)

For equal values of the original image to zero:

$Imagen_ecualizada(i,j) = histograma_acumulado(imagen_original(i,j) + 1)$ (4)

Where what goes in brackets is the index, or indexes, of every pixel of the image or of every value of the histogram. This way "original_image(i,j)" comes to indicate the index of the vector of the accumulated corresponding histogram. Hereby it is assigned to every pixel of the image equalized (or image with the histogram equalized) the density of accumulated probability corresponding to the value of the pixel of the original image. As this algorithm is done for Mat lab, and as Mat lab does not handle equal indexes to zero, it is considered that the histogram goes from 1 to 256, instead of 0 to 255. This way the density of probability of the value zero will be the one that is in the index 1 in the vector of the histogram.

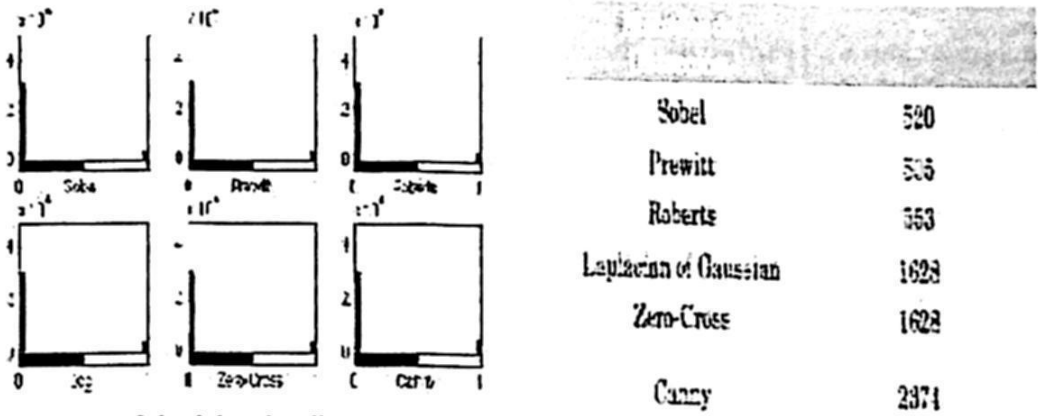


Fig. 8. Histogram of the blood cell recovers of six methods Edge

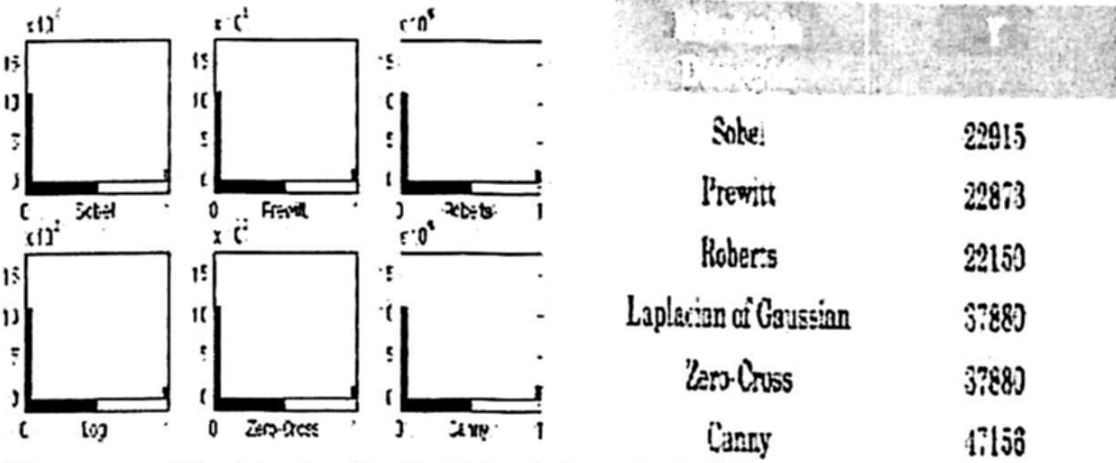


Fig. 9. Histogram of the blood cell with ALL of six methods Edge.

7 Morphologic operations

One of the operations mostly used in vision on images before binarizadas is the morphologic operations. The morphologic operations are operations realized on binary images based in forms. These operations take a binary image as an entry

returning an image also binary. The value of every pixel of the binary image is based on the value of the corresponding pixel of the original binary image and of its neighbors. Then choosing appropriately the form of the neighbors to consider, morphologic operations sensitive to a form can be constructed.

The principal morphologic operations are the expansion and the erosion. The operation of expansion adds pixels in the borders of the objects, while the erosion removes them. In both operations like I mention there is a grid used that determines neighboring, which of the central element of the grid will be born in mind for the determination of the proved pixel. The grid is a checkered arrangement that contains *some* and *zeros*, in the places that it contains *some* will be the neighbors of the original image with regard to the central pixel, which will be taken in consideration to determine the pixel of the image, whereas the places that have *zeros* will not be born in mind.

$$ImageR=erode (ImageS,w); \quad (5)$$

$$ImageR=dilate (ImageS,w); \quad (6)$$

The image shows graphically the effect of the grid on the original image and the result in the final image.

Since it shows the figure only of the yellow pixels in the original image take part in the determination of the red pixel of the image that is finally revealed.

Morphology

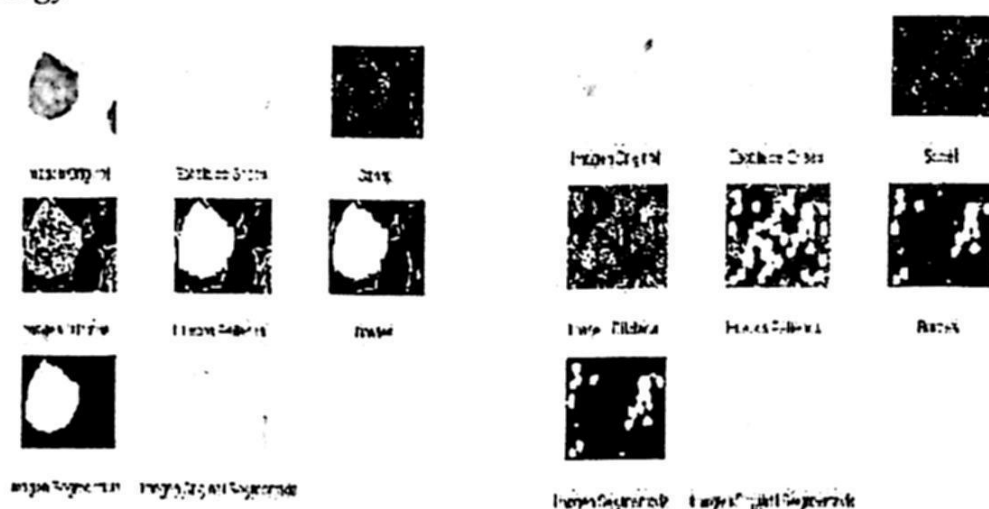


Fig. 9. Morphology: a) Healthy blood cell.

b) Blood cell with ALL



Fig. 10. Comparison Image Segmentation: a) Healthy blood cell. b) Blood cell with ALL

8 Characteristic

Table. 1. Scales for the detection of the ALL morphologic classification of the ALL

Characteri stic	L1	L2	L3	Mathematical Form
Cellular size	Small	Big	Big	Area or perimeter
Cromatina nuclear	Thin	Thin	Thin	Texture or histogram
Forms of the nucleus	Regular can have cracks	Irregular can have cracks	Irregular can have cracks	Circularity
Nucleus	Indistinguishable	Big Nucleus prominent	Big Nucleus prominent	Area or perimeter
Cytoplas m	Scanty	Abundant	Abundant	Area or perimeter

Some formulas that we can use:

- Nucleus and Cytoplasm area,

$$area = \sum_i \sum_j seg(i, j) \tag{7}$$

where $seg(i, j)$ are pixels of segmented object.

- Perimeter: The perimeter is the sum of the pixels of the contour of the object.
- Circularity:

$$Circularity = \frac{4 \cdot \pi \cdot area}{perimeter^2} \tag{8}$$

In our work we need an algorithm that helps us to classify the cells that are going to use the method of the diffuse logic. The diffuse logic is a technology of the computer like intelligence that allows to manipulate information with high degree of imprecision, it differs from the conventional logic that works with definite well and precise information.

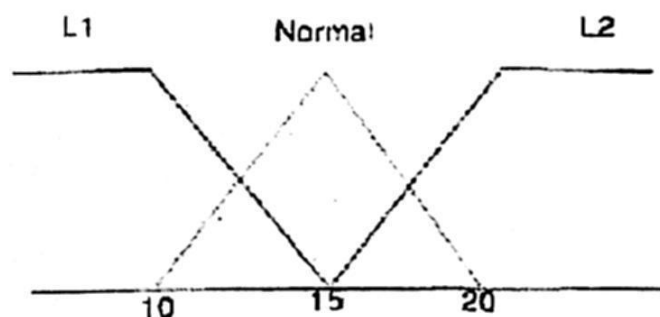
9 Fuzzy Logic.

WHAT IS THE Fuzzy LOGIC? The fuzzy logic is a methodology that provides a simple and elegant way of obtaining a conclusion from information of vague, ambiguous, vague entry that is incomplete, in general the fuzzy logic imitates how a persona takes decisions based on information with the mentioned characteristics. One of the advantages of the fuzzy logic is the possibility of implementing systems based on both hardware and software or in combination of both.

Normal Cell: 10-20 u diámetro.

L1: -20%

L2: +20%



10 Conclusion

We used two types of cells, a healthy cell and a cell with ALL. This way we could observe the differences between one and the other.

We considered six methods of segmentation for each of the cells, we also applied the morphology for each one.

We saw that the method Edge is not sufficient for the segmentation of the image with ALL, so I have decided to use another method (Histogram Equalization).

We are working on the extraction of characteristics of the cells this way we can implement the diffuse logic, the different diffuse sets are realized.

In the method of segmentation we use the function EDGE which has six methods: Sobel, Prewitt, Roberts, Laplacian of Gaussian, Zero-cross-country race and Canny.

We observed that the Canny method is the one that preserves more details, this way in the process of classification and recognition we are going to have a major certainty of analysis for it, I use the histogram of the images.

The part of the morphology is used to have a contrast of sufficient bottom; it is still necessary to improve some details to this part. we are still working on the extraction of characteristics of the cells, this way we can implement the Diffuse Logic, the different diffuse sets are realized.

Acknowledgements

The Instituto Politécnico Nacional for their support.

References

5. <http://www.cancer.gov/espanol/pdq/tratamiento/leucemia-linfoblastica-infantil/patient#Keypoint1>.
6. Harrison Principios de Medicina Interna.
7. Diagnóstico histoquímico de Leucemia Linfoblástica aguda infantil. Pérez – Chacón B., Ximena Gómez L., Patricia M. Sc. Instituto SELADIS Facultad de Cs. Farmacéutica y Bioquímica UMSA.
8. Guía de Referencia Rápida. Leucemia Linfoblástica Aguda 2009. Instituto Mexicano del Seguro Social. Dirección de Prestaciones Médicas. Unidad de Atención Médica. Coordinación de Unidades Médicas de Alta Especialidad. División de Excelencia Clínica.
9. Digital Imagen Processing whit Mat lab Programming, Rafael Gonzalez.
10. New Techniques in Oncologic Imaging, Anwar R. Padhani, Peter L. Choyke.
11. Feature Extraction & Image Processing, Mark Nixon & Alberto Aguado.

Formal Verification for the Absence of Deadlock in the Manager Workers Pattern

Jorge Luis Ortega-Arjona and Francisco Hernández-Quiroz

Facultad de Ciencias
Universidad Nacional Autónoma de México
Ciudad Universitaria, D.F. 04510, MEXICO
jloa,fhq@ciencias.unam.mx

Abstract. The Architectural Patterns for Parallel Programming are descriptions of the fundamental organizational features of common top-level coordinations observed in parallel software systems. They represent a means to capture and express experience in the design and development process of parallel software. Nevertheless, by now, these software patterns have been described in informal terms, in which very little can be stated about the properties present in the final parallel software system.

The present paper presents an initial approach for studying and documenting logical properties of an architectural pattern for parallel programming. In particular, the objective here is to formally verify the property known as “absence of deadlock” for the Manager-Workers pattern, a widely used architectural pattern for parallel programming, by means of formal verification using CCS and μ -calculus. The aim is to establish under what conditions this architectural pattern is deadlock-free, and whether this formal verification can be ported later to other Architectural Patterns for Parallel Programming.¹

Key words: Formal Verification, Absence of Deadlock, Manager-Workers pattern.

1 Introduction

Software patterns describe in a very general and abstract way a general problem in software design, and link it with a particular structure of software components that solve the general problem [3, 4]. Among all the software patterns, and in the area of parallel programming, the Architectural Patterns for Parallel Programming have been proposed as the *fundamental organizational descriptions of the common top-level structure observed in a group of parallel software systems* [8, 9]. They can be viewed as templates, expressing and specifying some structural properties of their communication and synchronization subsystems, and the responsibilities and relationships between them. The selection of an architectural

¹ This work was made possible thanks to the support of a project grant from our university (PAPIIT IN109010).

pattern for parallel programming is considered to be a fundamental decision during the design of the overall coordination of a parallel software system [9].

Architectural Patterns for Parallel Programming are defined and classified according to the requirements of order of data and operations, and the nature of their processing components. Requirements of order dictate the way in which parallel computation has to be performed, and therefore, impact on the software design [8, 9].

Nevertheless, even though these architectural patterns have served as guidance to the software designer or engineer, they still remain as a documented informal description about how to partition and communicate a problem among several parallel software components. In these terms, it would be advantageous to have further information about the performance and logic properties of the resulting parallel software system.

The objective of the present paper is to provide a formal verification that an important logical property of concurrency, namely the "absence of deadlock", is present in the Manager-Workers pattern (MW pattern hereafter), as an instance of an architectural pattern for parallel programming. For this, the MW pattern will be expressed as a CCS process [7] and absence of deadlock will be represented by a modal- μ calculus formula [5] satisfied by the process. The aim is to establish the conditions under which the MW pattern is deadlock-free, and if such a formal verification technique can be ported later to other Architectural Patterns for Parallel Programming. For our purposes here, deadlock is defined as the situation in which no process can take any further action but, at the same time, at least a process has a pending task [13].

[12] applied a similar approach to verification of mutual exclusion in parallel algorithms.

2 The Manager-workers pattern

The MW pattern is a variant of the Master-Slave pattern [3] for parallel systems, considering an activity parallelism approach where the same operations are performed on ordered data. The variation is based on the fact that components of this pattern are proactive rather than reactive. Each processing component simultaneously performs the same operations, independent of the processing activity of other components. An important feature is to preserve the order of data [8, 9].

The MW pattern has multiple data sets processed at the same time. So, a MW structure is composed of a manager component and a group of identical worker components. The manager is responsible of preserving the order of data. On the other hand, each worker is capable of performing the same independent computation on different pieces of data. It repeatedly seeks a task to perform, performs it and repeats; when no tasks remain, the program is finished. The

execution model is the same, independent of the number of workers (at least one). If tasks are distributed at run time, the structure is naturally load balanced: while a worker is busy with a heavy task, another may perform several shorter tasks. This distribution of tasks at runtime copes with the fact that data pieces may exhibit different size. To preserve data integrity, the manager program takes care of what part of the data has been operated on, and what remains to be computed by the workers [8,9].

2.1 Structure

The Manager-Workers pattern is represented as a manager, preserving the order of data and controlling a group of processing elements or workers. Usually, only one manager and several identical worker components simultaneously exist and process during the execution time. In this architectural pattern, the same operation is simultaneously applied in effect to different pieces of data by worker components. Conceptually, workers have access to different pieces of data. Operations in each worker component are independent of operations in other components.

The structure of this pattern involves a central manager that distributes data among workers by request. Therefore, the solution is presented as a centralized network, the manager being the central common component. An Object Diagram, representing the network of elements that follows the Manager-Workers structure is shown in Figure 1.

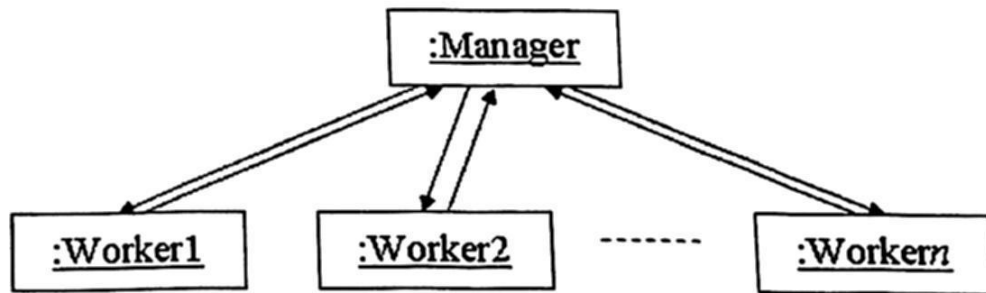


Fig. 1. Object Diagram of the Manager-Workers pattern.

2.2 Participants

- Manager. The responsibilities of a manager are to create a number of workers, to partition work among them, to start up their execution, and to compute the overall result from the sub-results from the workers.

- Worker. The responsibility of a worker is to seek for a task, to implement the computation in the form of a set of operations required, and to perform the computation.

2.3 Dynamics

A typical scenario to describe the run-time behavior of the Manager-Worker pattern is presented, where all participants are simultaneously active. Every worker performs the same operation on its available piece of data. As soon as it finishes processing, it returns a result to the manager, requiring more data. Communications are restricted between the manager and each worker. No communication between workers is allowed (Figure 2).

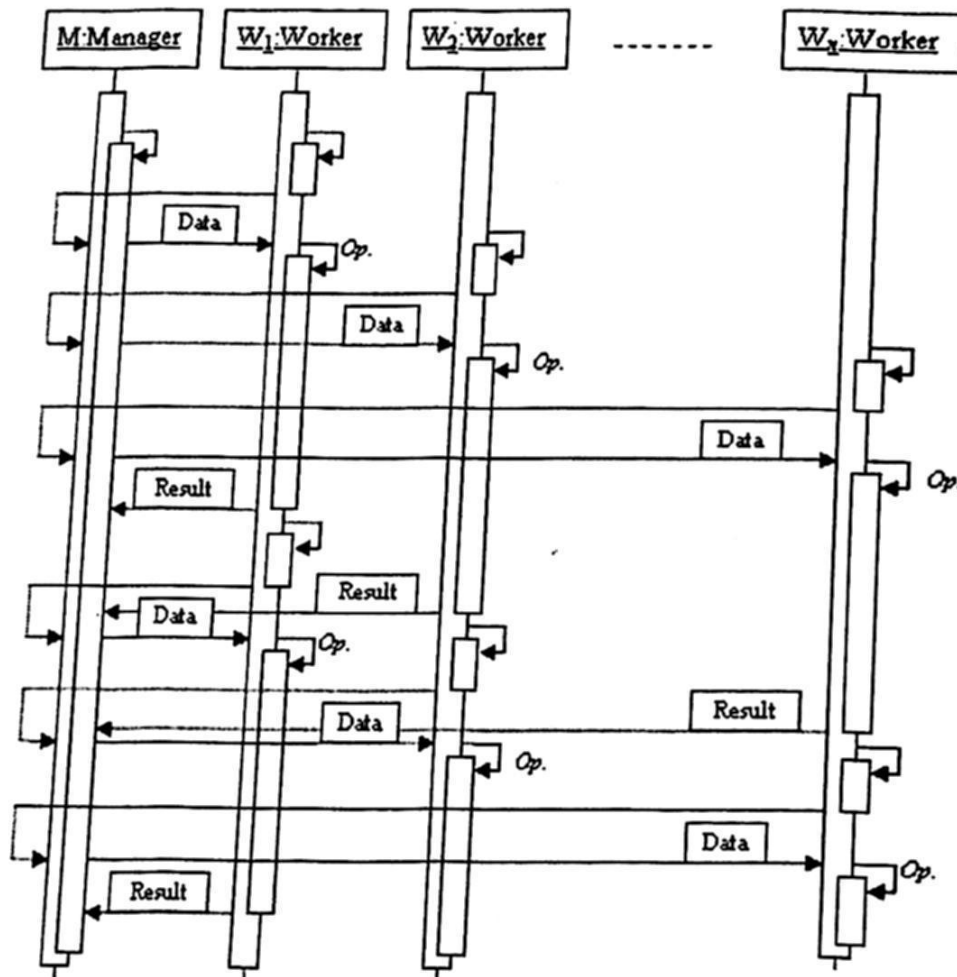


Fig. 2. Interaction Diagram of the Manager-Workers pattern.

In this scenario, the steps to perform a set of computations is as follows:

1. All participants are created, and wait until a computation is required to the manager. When data is available to the manager, this divides it, sending data pieces by request to each waiting worker.
2. Each worker receives the data and starts processing an operation $Op.$ on it. This operation is independent of the operations on other workers. When the worker finishes processing, it returns a result to the manager, and then, requests for more data. If there is still data to be operated, the process repeats.
3. The manager is usually replying to requests of data from the workers or receiving their partial results. Once all data pieces have been processed, the manager assembles a total result from the partial results and the program finishes. The non-serviced requests of data from the workers are ignored.

3 CCS

Milner designed the Calculus of Communicating Systems [7] for modelling concurrency in systems that communicate in a message-passing synchronous style. Messages are sent via two-end channels. Access to channels can be *open* or *restricted*. A message is consumed once communication has taken place and therefore is no longer available to anybody else.

Formally speaking, we have a countable set of channels $\alpha, \beta, \gamma, \alpha_0, \dots$. The basic actions are

- sending a message: $\alpha!m$, where α is a channel;
- receiving a message: $\alpha?m$;
- synchronous communication between process without any disclosure to outsiders: τ .

Let us denote by the variables λ, λ' any of the first two types of actions. If λ is the action $\alpha?m$, then $\bar{\lambda}$ is the complementary action $\alpha!m$ and viceversa.

Processes are made out of basic actions and the simple process *nil* compounded by *prefixing* by basic actions, *non-deterministic choice*, *parallel composition*, *restriction* of channels and *relabelling* of channels. In BNF:

$$p ::= nil \mid \lambda . p \mid (p + p) \mid (p \parallel p) \mid p \backslash L \mid p[f],$$

where L is a set of channels and f is a one-to-one mapping of channels.

Finally, a process can be defined recursively by an equation

$$N \equiv_{\text{def}} p$$

where N is a name and P is a process in the above grammar (possibly containing instances of the name N).

The semantics of CCS is usually expressed in terms of structural operational semantic rules [11] with labelled transitions:

Prefixed process

$$\lambda . p \xrightarrow{\lambda} p$$

Choice

$$\frac{p \xrightarrow{\lambda} q}{(p + r) \xrightarrow{\lambda} q} \quad \frac{r \xrightarrow{\lambda} q}{(p + r) \xrightarrow{\lambda} q}$$

Parallel composition

$$\frac{p_0 \xrightarrow{\lambda} p'_0}{p_0 \parallel p_1 \xrightarrow{\lambda} p'_0 \parallel p_1} \quad \frac{p_1 \xrightarrow{\lambda} p'_1}{p_0 \parallel p_1 \xrightarrow{\lambda} p_0 \parallel p'_1} \quad \frac{p_0 \xrightarrow{\lambda} p'_0 \quad p_1 \xrightarrow{\bar{\lambda}} p'_1}{p_0 \parallel p_1 \xrightarrow{\tau} p'_0 \parallel p'_1}$$

Channel restriction

$$\frac{p \xrightarrow{\lambda} q}{p \setminus L \xrightarrow{\lambda} q \setminus L} \quad \lambda \notin L \cup \bar{L}$$

Channel relabelling

$$\frac{p \xrightarrow{\lambda} q}{p[f] \xrightarrow{f(\lambda)} q[f]}$$

Recursively defined processes

$$\frac{p \xrightarrow{\lambda} q}{P \xrightarrow{\lambda} q} \quad \text{where } P \equiv_{\text{def}} p.$$

3.1 Manager-Workers in CCS

Our translation of the Manager-Workers pattern to CCS focuses on the messages sent from the manager to assign a task, and the results sent back by the worker. The specific task performed for the worker can be left unspecified.

The components of the main process (represented by M) will be as follows: (a) a task being assigned (represented by T); (b) the worker which receives the assignment (represented by W); (c) the assignment itself (represented by A). Subscripts will be used to distinguish between different workers and tasks.

$$\begin{aligned} W_i &\equiv_{\text{def}} (\alpha_i ? m . P . \alpha ! m . \text{nil}) \\ T_i &\equiv_{\text{def}} (\alpha_i ! m . \alpha ? m . A_i) \\ A_i &\equiv_{\text{def}} (T_i \parallel W_i) \setminus \{\alpha_i\} + \text{nil} \\ M &\equiv_{\text{def}} C \parallel (A_1 \parallel \dots \parallel A_n) \end{aligned}$$

A further explanation is needed:

- Worker W_i is expecting through channel α_i an assignment. When this happens, it proceeds to perform the task, which is represented by the unspecified process P . The only condition we will assume later is that P is deadlock free itself. For this assumption to be realistic, P should not depend on any external synchronizing event. Then the result is sent back through the same channel.
- Task T_i contains the complementary communication actions of W_i . Once a task is finished control is handed over to the assignment process.
- Assignment A_i can choose either calling in a task *and* a worker to perform it, or consider the work done and become *nil*. Observe how channel α_i is restricted in order to guarantee integrity of the communication with W_i .
- M is the manager creating tasks and combining results in process C , while in parallel assigns tasks to different workers. Please do note that C is also left unspecified and again the only condition required is that C is deadlock free.

A final but important remark: this translation has made explicit important facts about synchronization between actions from manager and workers which by no means were stated in the English description of this pattern and that will be critical to have a deadlock free system (as it will be proved later).

4 Model Checking

For expressing properties of processes we will be using a version of modal μ -calculus extended with special constants [5]. μ -calculus is a propositional multi-modal logic with an additional least-fixed point operator for recursive formulas. The modal propositional part of the language is essentially Hennessy-Milner logic [6].

μ -calculus has as atomic propositions the logical constants V y F (we will add some more atomic propositions later). The modalities in HML are labelled by CCS basic actions. In BNF:

$$D ::= V \mid F \mid \neg D \mid D \vee D \mid D \wedge D \mid \langle \lambda \rangle D \mid \langle \cdot \rangle D \mid \mu X . D$$

We can add the following abbreviations:

$$[\lambda]D \equiv_{\text{def}} \neg \langle \lambda \rangle \neg D \quad [\cdot]D \equiv_{\text{def}} \neg \langle \cdot \rangle \neg D$$

We define inductively the satisfaction relation \models between CCS processes and HML formulas:

$$\begin{aligned}
 p &\models V && \forall p \in \text{CCS} \\
 p &\not\models F && \forall p \in \text{CCS} \\
 p &\models \langle \lambda \rangle D && \text{iff } \exists p' . p \xrightarrow{\lambda} p' \wedge p' \models D \\
 p &\models \langle \cdot \rangle D && \text{iff } \exists p' . \lambda . p \xrightarrow{\lambda} p' \wedge p' \models D \\
 p &\models [\lambda] D && \text{iff } \forall p' . p \xrightarrow{\lambda} p' \Rightarrow p' \models D \\
 p &\models [\cdot] D && \text{iff } \forall p' . \lambda . p \xrightarrow{\lambda} p' \Rightarrow p' \models D \\
 p &\models \mu X . D && \text{iff } p \models D_{[X := \mu X . D]}
 \end{aligned}$$

For instance, the process $(\alpha?m . nil) \parallel (\beta!n . nil)$ satisfies formula $\langle \alpha?m \rangle [\cdot] V$ because

$$(\alpha?m . nil) \parallel (\beta!n . nil) \xrightarrow{\alpha?m} nil \parallel (\beta!n . nil)$$

and the latter process can perform only one action and therefore the only possible transition is

$$nil \parallel (\beta!n . nil) \xrightarrow{\beta!n} nil \parallel nil$$

and $nil \parallel nil \models V$. On the other hand, the same process does not satisfy $[\cdot] \langle \alpha?m \rangle V$, for if it performs firstly the transtion (and we are obliged to take into account every possibility by the operator $[\cdot]$)

$$(\alpha?m . nil) \parallel (\beta!n . nil) \xrightarrow{\alpha?m} nil \parallel (\beta!n . nil)$$

we will have

$$nil \parallel (\beta!n . nil) \not\models \langle \alpha?m \rangle V.$$

4.1 Formal properties

The other side of model checking verification requires to express the desired/undesired properties as formulas in a logical language. In this case, we need to express deadlock as a HML formula. Following [13], we introduce a new atomic formula satisfied by processes with no pending tasks.

$$\begin{aligned}
 nil &\models \text{terminal} \\
 \lambda . p &\not\models \text{terminal} \\
 p + q &\models \text{terminal} \text{ if } p \models \text{terminal} \text{ and } q \models \text{terminal} \\
 p + q &\not\models \text{terminal} \text{ otherwise} \\
 p \parallel q &\models \text{terminal} \text{ if } p \models \text{terminal} \wedge q \models \text{terminal} \\
 p \setminus L &\models \text{terminal} \text{ if } p \models \text{terminal} \\
 p[f] &\models \text{terminal} \text{ if } p \models \text{terminal}
 \end{aligned}$$

Our working deadlock definition implies that there is no possible action that can be performed and yet the process still has pending tasks. For the first part, a process capable of no action will satisfy trivially the formula $[\cdot] F$. But if the

process has not terminated yet it will not satisfy *terminal*. Then deadlock can be defined by

$$dead \equiv_{\text{def}} ([\cdot] F \wedge \neg terminal).$$

Nevertheless, this formula represents the fact of the process that already has reached deadlock. We want to say that a process can or cannot deadlock now or in the future (ie. after performing a certain number of actions). For this, we need the recursive formula

$$e\text{-}dead \equiv_{\text{def}} \mu X . (dead \vee \langle \cdot \rangle X).$$

A process satisfying this formula may eventually deadlock. In the following we will check our translation of the MW pattern.

4.2 Deadlock absence for MW

Let us see now where our Manager-Workers representations stand regarding eventual deadlock. We need to answer

$$M \models e\text{-}dead?$$

According to the rules for recursive formulas, this is equivalent to

$$M \models dead \vee \langle \cdot \rangle e\text{-}dead?$$

Being a disjunction we need both disjuncts. Let us start with *dead*, that is

$$M \models [\cdot] F \wedge \neg terminal?$$

Assuming *C* is deadlock free we can focus on the component $(A_1 \parallel \dots \parallel A_n)$ and check whether

$$(A_1 \parallel \dots \parallel A_n) \models [\cdot] F \wedge \neg terminal.$$

We have two cases: (a) *nil* is chosen in each instance of A_i and we end up trivially with a collection of *nil*; (b) at least one of the assignments launches an instance of $(T_i \parallel W_i)$.

In case (a), $(nil \parallel \dots \parallel nil)$ is equivalent to *nil* and although $nil \models [\cdot] F$, by definition $nil \not\models terminal$ and then $nil \not\models dead$.

In case (b), by virtue of the rules of structural operational semantics

$$(T_i \parallel W_i) \xrightarrow{\tau} (\alpha?m . A_i) \parallel (P . \alpha!m . nil)$$

which means that

$$\begin{aligned} (A_1 \parallel \dots \parallel (T_i \parallel W_i) \parallel A_n) &\not\models [\cdot] F && \text{and} \\ (A_1 \parallel \dots \parallel (T_i \parallel W_i) \parallel A_n) &\not\models \neg terminal \end{aligned}$$

and therefore

$$M \not\models dead.$$

What about the other side of the disjunction, namely $\langle \cdot \rangle e\text{-dead}$? Applying again the rule for recursive formulas, we are asking the question

$$M \models \langle \cdot \rangle \text{dead} \vee \langle \cdot \rangle \langle \cdot \rangle e\text{-dead?}$$

Let us consider again the first disjunct. As before, we are faced with the question

$$(A_1 \parallel \dots \parallel A_n) \models \langle \cdot \rangle \text{dead},$$

and again we can have the two cases (a) and (b). In case (a)

$$(nil \parallel \dots \parallel nil) \not\models \langle \cdot \rangle \text{dead},$$

this time by definition of \models for the operator $\langle \cdot \rangle$ (as there is no possible action). Suppose now that a process A_i chooses to launch an instance of $(T_i \parallel W_i)$. Following a similar argument as before, we have to answer now the question

$$(A_1 \parallel \dots \parallel (T_i \parallel W_i) \parallel A_n) \models \langle \cdot \rangle \text{dead?}$$

Again, we apply the rules of structural operational semantics

$$\frac{(T_i \parallel W_i) \xrightarrow{\tau} (\alpha?m . A_i) \parallel (P . \alpha!m . nil)}{(A_1 \parallel \dots \parallel (T_i \parallel W_i) \parallel A_n) \xrightarrow{\tau} (A_1 \parallel \dots \parallel ((\alpha?m . A_i) \parallel (P . \alpha!m . nil)) \parallel A_n)}$$

and the rules of \models pose now the question

$$(A_1 \parallel \dots \parallel ((\alpha?m . A_i) \parallel (P . \alpha!m . nil)) \parallel A_n) \models \text{dead?}$$

Given that P is deadlock-free and does not require any external communication event we now that

$$\begin{aligned} (\alpha?m . A_i) \parallel (P . \alpha!m) &\xrightarrow{\tau} *(\alpha?m . A_i) \parallel (\alpha!m . nil) \text{ and} \\ (\alpha?m . A_i) \parallel (\alpha!m . nil) &\xrightarrow{\tau} A_i \parallel nil \end{aligned}$$

ie, the process can perform at least one action and therefore, as before,

$$(A_1 \parallel \dots \parallel ((\alpha?m . A_i) \parallel (P . \alpha!m)) \parallel A_n) \not\models \text{dead}.$$

This again resolves the left-hand side of the second disjunction, and we have to look at $\langle \cdot \rangle \langle \cdot \rangle \text{dead}$ which is tantamount to

$$M \models \langle \cdot \rangle \langle \cdot \rangle \text{dead} \vee \langle \cdot \rangle \langle \cdot \rangle \langle \cdot \rangle e\text{-dead}.$$

Following the previous line of reasoning we will be considering either the question

$$(A_1 \parallel \dots \parallel (A_i \parallel nil) \parallel A_n) \models \text{dead?}$$

which we have already answer on the negative.

But again we will have to consider a new right-hand side disjunct. This can lead us to think this process will never end. However, it is not difficult to see by now that the new questions posed can be reduced to one of the previously considered, no matter how many times a task and its corresponding worker is launched. Therefore it is safe to state that

$$M \not\models e\text{-dead}.$$

5 Conclusions and Future Work

We have shown a way of translating a natural language description of the MW pattern into a syntactic construction of a formal language. This translation helped to expose and solve some vagueness implicit in the original description and it also made explicit decisions of synchronization previously left to application phase of the pattern.

The new formulation of the MW pattern was proved to comply with the property of deadlock absence. For this we used (an instance of) the modal μ -calculus and standard model checking techniques. So we can ascertain that a system based on this formulation of the MW pattern will be deadlock-free (provided the additional assumptions stated here are also met).

The MW pattern is a very general one and it can be specialized into more particular patterns according to specific decisions related to coordination order, discipline of communication or any other consideration deemed relevant. Future work should consider these variants in order to prove that they are also deadlock free.

Absence of deadlock can be defined differently, as in [1, 2]. Different definitions may require different formulas in μ -calculus and either a proof of equivalence or inclusion between them or separate proofs of deadlock absence.

Finally, we hope we have shown the utility of this method and how it can be applied to other Architectural Patterns for Parallel Programming.

References

1. Andrews, G.R., *Concurrent Programming*, The Benjamin/Cummings Publishing Company, 1991.
2. Andrews, G.R., *Foundations of Multithreaded, Parallel, and Distributed Programming*, Addison-Wesley, 2000.
3. Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., Stal, M., *Pattern-Oriented Software Architecture: A System of Patterns*, John Wiley & Sons, 1996.
4. Gamma, E., Helm, R., Johnson, R., and Vlissides, J., *Design Patterns: Elements of Reusable Object-Oriented Systems*, Addison-Wesley, 1994.
5. Kozen, D., "Results on the propositional μ -calculus", *Theoretical Computer Science* 27, pp. 333-354, 1983.
6. Larsen, K.G., "Proof systems for Hennessy-Milner logic with Recursion", *Lecture Notes In Computer Science*, 299, Proceedings of the 13th Colloquium on Trees in Algebra and Programming, pp. 215-230, 1988.
7. Milner, A.J.R.G., *Communication and Concurrency*, Prentice Hall, 1989.
8. Ortega-Arjona, J.L., *Architectural Patterns for Parallel Programming. Models for Performance Estimation*, VDM Verlag, 2009.
9. Ortega-Arjona, J.L., *Patterns for Parallel Software Design*, John Wiley & Sons, 2010.
10. Najm, E., Stefani, J.-B. *Formal Methods for Open Object-based Distributed Systems*, International Workshop 1996, Paris, Chapman & Hall, 1997.
11. Plotkin, G.D., *A structural approach to operational semantics*, DAIMI FN-19, Computer Science Department, Aarhus University, 1981.

12. Walker, J., "Automated analysis of mutual exclusion algorithms using CCS", *Formal Aspects of Computing*, 1:1, pp. 273-292, 1989.
13. Winskel, G., "A note on model checking the modal mu-calculus", *Theoretical Computer Science* 83, pp. 761-772, 1991.

Medical Carnet for Management of Patients Driven by Ontologies

F. Mata and S. Zepeda

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas,
Instituto Politécnico Nacional.
Avenida Instituto Politécnico Nacional No. 2580, Colonia Barrio la Laguna
Ticomán, Delegación Gustavo A. Madero, CP. 07340 México D.F.
mmatar@ipn.mx, pcortez@ipn.mx, sazepeadar@gmail.com

Abstract. Nowadays, management of patients in hospitals and clinics is automated by computer systems. Nevertheless, the patients no have a direct participation in this process. If patients would have a digital medic carnet, stored in his/her cell phone, other clinics and hospitals could explore this information, facilitating diagnosis and medical treatments. We propose a web and mobile system based on digital medic carnet stored into cell phone. The medic carnet will be available for doctors and patients using Bluetooth communication from a computer or cell phone. Ontology is designed to assist doctors in the diagnosis and treatment of diseases, and to assign a specialist according to ailment and previous diagnosis. The Medical carnet represents the clinic history of patient, it will be available for paramedics, and will be very useful in emergency cases, when is required to provide immediate medical attention. Then, ontology proposes treatments according to possible allergies and sufferings of the patient. The administration and management of the patient's medical history is done through a web application. Were tested Bluetooth communication among cell phones, cell phones-computer and computers-computers. The main contribution is the integration of semantic similarity in a mobile and web application oriented to doctors and patients.

Keywords: Mobile software, Ontology, Management patients, Bluetooth communication.

1 Introduction

Presently, Bluetooth technology has become the most popular media for sharing information between mobile devices, cell phones, computers and other software applications. In addition, the mobile software has been consolidated, including applications addressed not only to communicate to people, but also for entertainment, assistance, electronic commerce and to store personal information of user (even private and confidential). In this direction, the documents of common use, such as identification and medical carnets which contains medical history of user and affiliation to a hospital. The information contained in this document is available only

when people visit doctor or hospital. Therefore, a mobile software for management a medical carnet, would be easier achieve tasks of diagnosis, treatment and monitoring in case of accidents, where information can be accessed even when the patient is unable to provide this information to paramedics, because serious psychological trauma that causes an accident.

In addition, software installed into cell phone that be the equivalent in use and usefulness of the paper medical carnet represents several benefits. In this sense, according to INEGI, 61% of the families in Mexico already have cell phones [1]. Moreover, considering ontologies development and semantic processing let to offer solutions, suggestions, and even assist into making decisions in the medical field and hospitals. From diagnosis to suggested treatments according to patient's medical history. The system presented uses the capabilities and technologies for mobile devices to create an electronic medical carnet that is stored on mobile devices, and can be accessed via Bluetooth, and Web by a paramedic, in case of an emergency. System offers suggestions of doctors who can treat a patient and other treatments that can be applied to a patient.

This is the main motivation of our work, the integration of wireless communications, mobile software, representations of knowledge such as ontologies, and the consolidation of the web to provide a solution for patients and doctors in clinics and hospitals in tasks of diagnosis and treatment.

2 Problem statement

Today, medics and paramedics face several problems, one of them is to provide adequate care when no information about the patient is available, in cases where a patient suffers an accident or he/she cannot provide this information. The protocol indicates that before treating the patient, the paramedic must make a series of questions to offer a diagnosis of patient condition and to know previous suffering or medical conditions that may require special treatment. But at least two cases are not considered: when patient does not know all his/her medical data, then it results that information can be wrong or incomplete, the second one occurs in accidents, where persons loses the ability to communicate, in this case, paramedics are forced to work on uncertainty, risking the life of person: any decision of paramedic could be counterproductive for the patient. Therefore, in these cases, a system would be useful, if this provides the information necessary to offer a safe and effective treatment to the patient; this is part of our proposal.

3 State of art

The development of mobile applications addressed to health care has started with other systems, such as "Primary care plus ambulatory and Hospital Care Suite (PCP)" [2] which is a suite that let to see medical information on Palm OS, Windows Mobile, or Internet. There is also the case "Emedic" is a medical encyclopedia. While in [3]

people who have a 3G mobile device may have their medical records (including x-rays, records of pressure, etc.) disadvantage is the cost of these devices and the cost of mobile application.

The communication system "Vocera" [4], developed in St.Vincent hospital, Birmingham, USA. It is a communicator badge system for mobile users with a push-to-call button, a small text screen and voice-dialing capabilities based on speech recognition. Moreover, the "Context-aware mobile communication" [5], developed in Mexico provide contextual messaging, for example, "a message for room 226 to any doctor, delivery time for the message today after 2 p.m.". while that "Intelligent hospital software" [6] provide remote query, tracking of patients and equipment, notification of awareness and patient data, an experimental prototype is implemented with function for starting an audio-video conference from the nearest point. Doctors are localized, are notified of the call. This prototype is presented as a demonstrator of the middleware platform QoS DREAM, for reconfigurable multimedia streaming and event-based programming. The first two applications only serve as reference for the diagnosis; the third has the patient information, however, the disadvantage that mobile devices should be 3G. In contrast our system can be used in any system that supports Java J2ME technology and JSR82 specification.

4 Methodology

Our methodology is a system composed of three applications: 1) A mobile application for patients and doctors 2) A web application to medical management in hospitals 3) an ontological module for assisting doctors in diagnosis tasks and treating diseases.

The former has the ability to store, send and receive medical cards via Bluetooth. Later, the medic carnets will be available by Web application. An ontology that stores concepts: patients, doctors, diseases, diagnosis and treatment received by patient. The system architecture consists of three modules, the module that resides in the hospital where administrators, paramedics and patients are registered. In addition, the ID card of paramedic and medical carnet of patient are transferred to their respective cellular phones. In addition, save reports generated on the paramedic cell phone, these are received by the hospital terminal via Bluetooth. The second one and third module are mobile applications that are installed in cell phones of paramedics and patients respectively. These are shown in Figure 1.

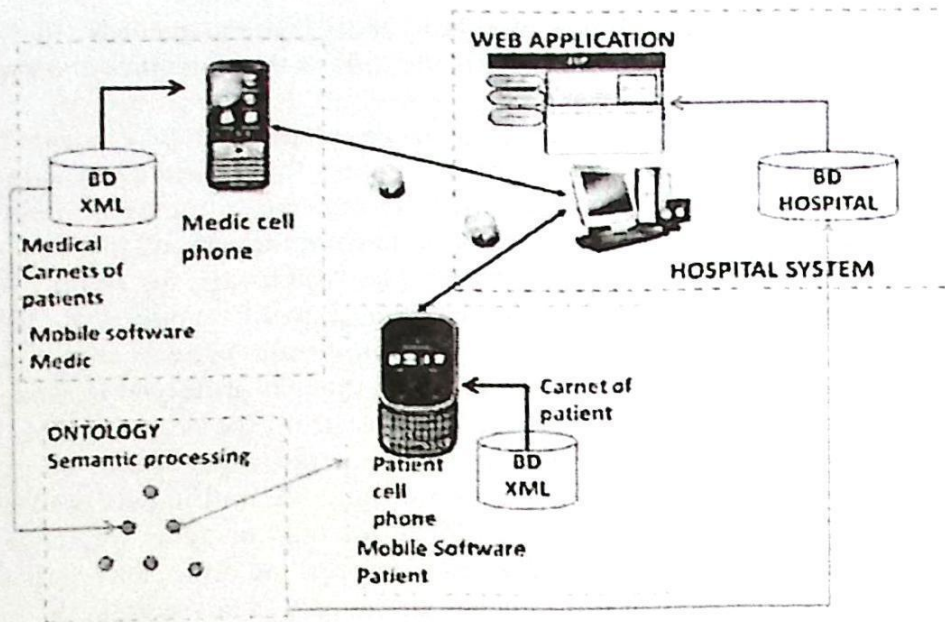


Fig. 1. Framework of system

As shown in Figure 1 we have three modules: the mobile module for patient and doctor, the web module and the module ontological. The web system module manages user information, classifying them as administrators, paramedics and patients. Registration of a paramedic or a patient is achieved by generating an XML ID card which is sent to cell phone of patient. When a modification on card occurs, then administrator can update the carnet stored into his/her cell phone. The process of sending and receiving data by Bluetooth is made through the mobile module. Paramedics make reports, starting session by sending ID card by Bluetooth, then patient files can be transferred, as well as medical records, events attended and care provided to patients.

The mobile module requires the establishment of Bluetooth communication search for services is performed in hidden mode for mobile phones and not in other devices to reduce the time of connection establishment. In this scenario to resolve the potential conflict posed by the existence of several cell phones that have the same application surround of paramedics, then, only the paramedic will activate the application manager on cell phone of patient, by sending a password. It provides confidentiality of the information and ensures that only will be obtained the patient's medical history data.

The ontology module processes semantically the information contained into mobile database and from web database. The ontology contains concepts related to patients and their ailments, and relationships among patients and doctors according to certain disease. The semantic processing of these concepts and relationships let offer two functions: 1) suggestions in the medical diagnostics (processing the semantic similarity of medical treatment), 2) suggest a doctor or specialist according to diagnosis that give the doctor (similarity in diagnosis). The ontology structure is show in Figure 2.

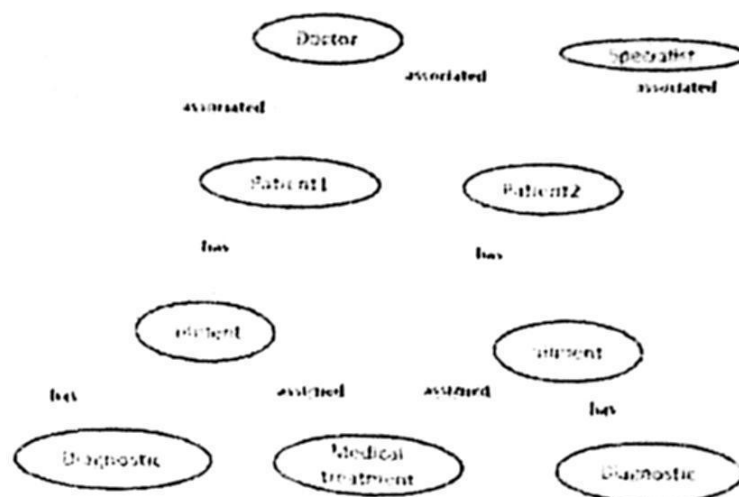


Fig. 2. Conceptual structure of Ontology.

The ontology was implemented in OWL and is explored using the SPARQL language, and as shown in Figure 2, have six concepts are: Patient, Doctor, Specialist, ailment, Medical diagnostic and treatment. Through the semantic relationships between concepts (has and assigned) you can find doctors and specialists who have examined a patient, what treatments have been applied, and sufferings undergone by the patient. Finding similarities among treatments, and offer suggestions in the medical diagnosis, and suggest a doctor or specialist according to medical historic.

Semantic processing is performed in two steps: 1) exploring the ontology to find the requested concept, 2) Extracting context and semantic relationships among concepts are processed.

To explain these steps, we consider the ontological module functions, where the first one is to determine the semantic similarity in medical treatments, for offering suggestions in diagnosis tasks.

For the first step, suppose we search: "ailment", ontology is explored from the parent node until you find the concept associated to "ailment". When this is found, second step is applied, which extracts the context of the concept searched, in this case is ailment (the context are the concepts associated by relationships "has" and "assigned"). Then, assigned relationship lets know what medical treatments have been applied to other patients. While the semantic relationship "has" lets to get other patients that have suffered the same disease. Then you can compare the treatments applied to other patients to find semantic similarities between them (e.g. the chemical substance of different medicaments used in treatments). Moreover, the system indicates when several patients have been reported with the same disease in a short time range (days or hours), indicating that it may be an epidemic.

Now, we explain the semantic processing for the second function of the ontological module, which is the similarity in diagnosis, e.g. to recommend a doctor or specialist according to diagnosis evaluated by another doctor. This process is to know which doctors or specialists have treated several patients. Therefore, we explore the ontology to find the concept associated with a doctor, if it is found, patients associated with that doctor are extracted, and if exists the specialists that have examined the patient. We

compared the diagnosis given by each doctor to each one of these patients and if exist a similarity suggests which specialist or doctor is the indicated.

The system displays in graphical form the associations among these concepts to assist in decision making by medical staff and view description of each instance associated with the concept (e.g. Dr. Perez associated with the concept Doctor). In the next section the results are shown.

5 Experiments and Results

The experiments were performed using cell phones Sony Ericsson w580 and Motorola 810. According to the following roles of user: patients, doctors or paramedic. We started by explaining the functions addressed to patients for mobile and web. The first test was to register a user (administrator, paramedic or patient) for them to logon and typing the required data into a web form, as shown in Figure 3.

The screenshot shows a web browser window with a registration form titled "Registro de Administradores". The form contains several fields for user information, including name, email, password, and role. A sidebar on the left lists navigation options like "Inicio", "Registro de Usuarios", and "Consulta de Datos".

Registro de Administradores	
Nombre:	<input type="text"/>
Apellido:	<input type="text"/>
Correo Electrónico:	<input type="text"/>
Contraseña:	<input type="password"/>
Confirmar Contraseña:	<input type="password"/>
Seleccionar Rol:	<input type="text"/>
<input type="button" value="Registrar"/>	

Fig. 3. User registration on Web

After user has been registered a identification card (ID) is generated and it is sent by Bluetooth to cell phone of user, the result of this operation is shown in Figure 4.

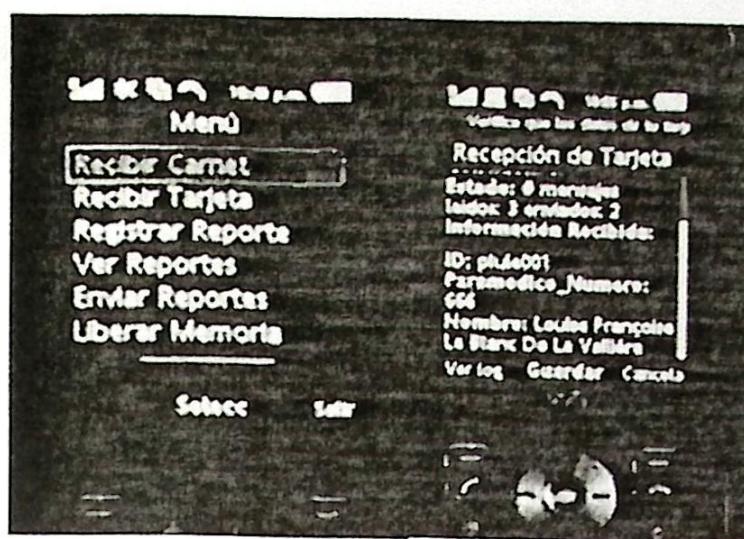


Fig. 4. Medical carnet reception

As shown in Figure 2, data from the medical card is received and displayed on the user's cell phone. Now we explain the testing for role of user: doctors. In this case the patient's medical card is explored from a Web form where the doctor can know the various treatments that have been applied to patients for a specific disease. This query is not processed from a database, but is performed through the SPARQL language using the concepts stored in the ontology. The difference is that the concepts are structured for performance the role of a medical assistant. The results displayed will facilitate the decision making according to treatment assigned to patients. The web interface of this process is shown in Figure 5.

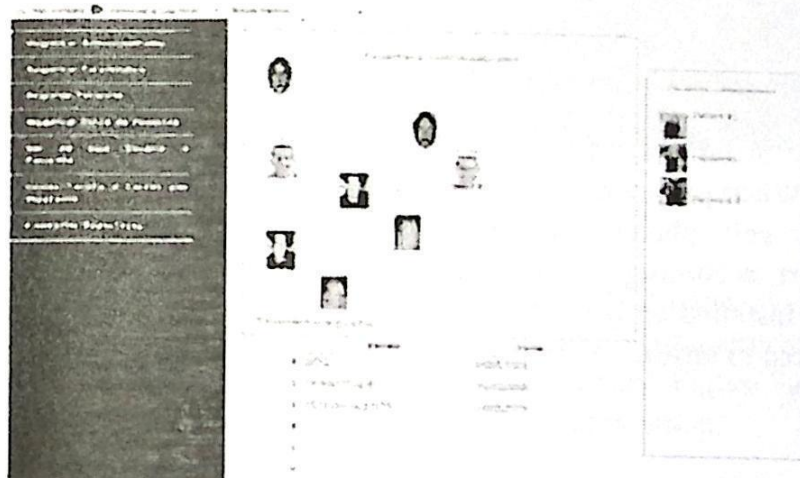


Fig. 5. Similarity by medical treatment

As shown in Figure 5, the relationship between patients and doctors are shown, and when a patient is selected, the various medical treatment received (medical history) are displayed, moreover, shows the related patients, e.g. patients with diseases in common and have been treated by the same or doctors or specialists. The treatments are compared conceptually (semantic similarity) For example, medicaments can be supplied with different names but they containing the same chemical substance) that

can find out by semantic similarity, where the ontology is assisted by a taxonomy of medicaments and chemical substances. The concepts of such medicaments have instances (chemicals substances). Thus, according to the degree of similarity between medicaments and treatments provided to assigned patients. The system gives suggestions on treatments for a particular disease.

Another function offered by the system is to suggest a specialist or physician for a particular disease, according to the diagnosis given by doctor who reviewed. This compares assigned diagnosis and treatments for a disease, in different patients. Figure 6 shows the interface for this process.

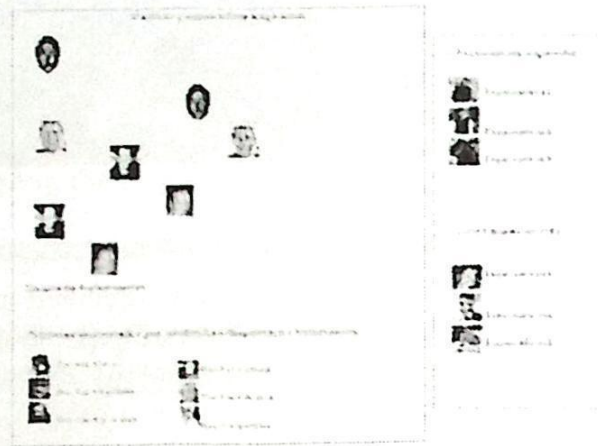


Fig. 6. Suggestion of a medico or specialist

The comparison of diagnoses and treatments were performed by identifying the same disease in different patients, compare the concepts representing the treatment and diagnosis, if a matching is found the doctors or specialists associated with them are suggested as candidates appropriate to treat the patient.

Moreover, the patient registration needs to provide medical information to be stored into medical carnet. Personal data are entered and information about diseases and allergies of patient. After patient registration the carnet is generated and sent to his/her cell phone. When paramedic is on a emergency, he establishes communication with the patient's cell phone and access to patient's carnet, generating the corresponding report, according to vital signs, treatments provided, and the material used. Applications installed on the cell phones of the paramedic and patient using a security code required to ensure safety in the transfer of confidential information. This is shown in Figure 7.

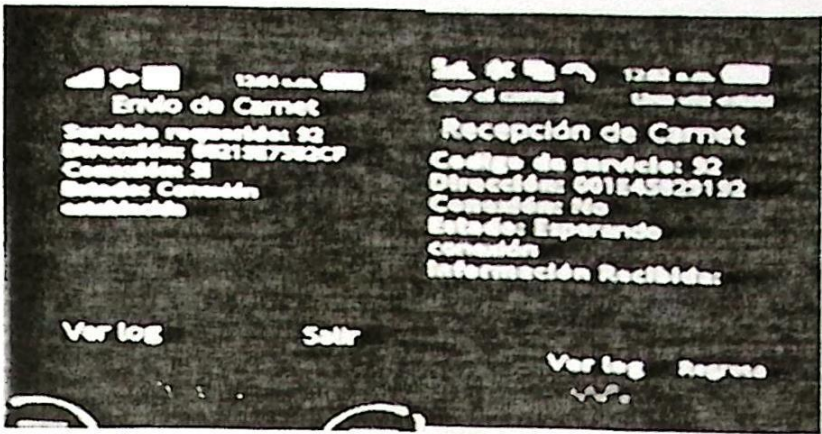


Fig. 7. Carnet transferring

Furthermore, reports include register vital signs such as temperature and pressure. Additionally, the medicines and materials used in the event. To complete the report paramedics provides the location data of the event. This is shown in Figure 8.

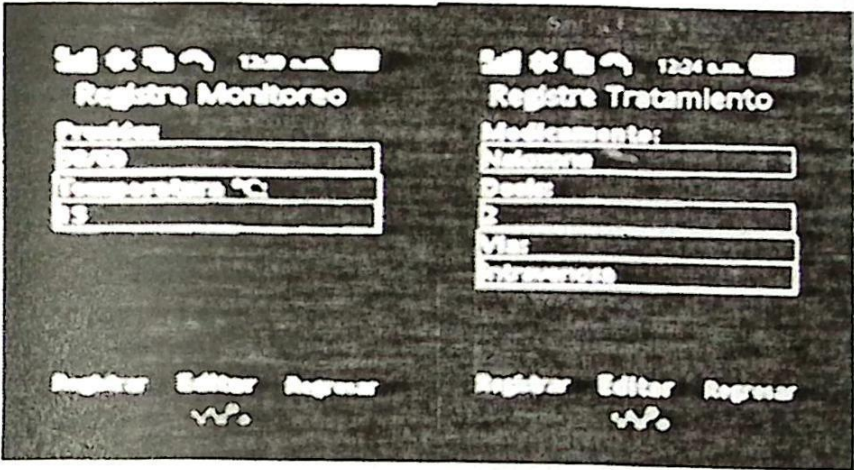


Fig. 8. Report view

In addition, paramedics can enter their reports to the system and to query previous reports. The session is started with Bluetooth communication by sending the digital ID, and SSL protocol to increase system security. Figure 7 shows the result of the transfer and registration of a report in the web application.

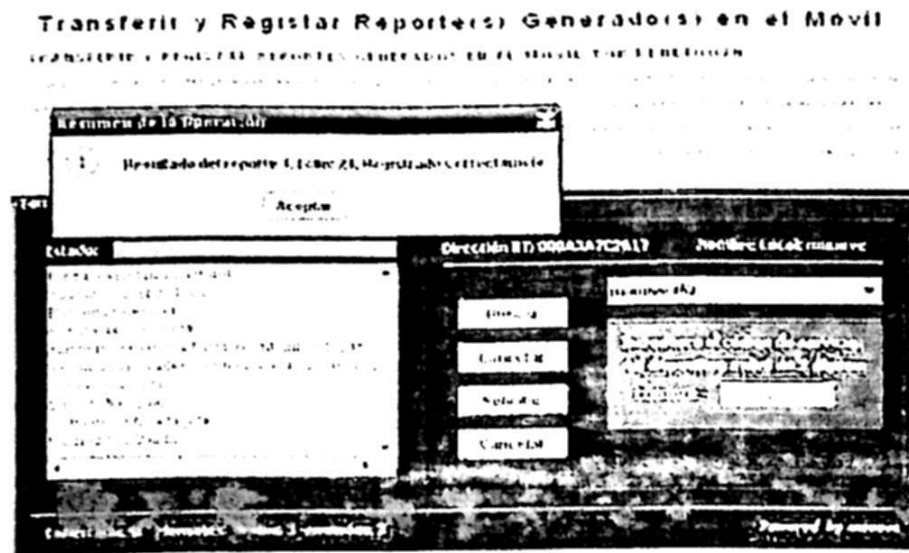


Fig. 9. Report transfer

6 Conclusions

We present a system to manage the patient's medical history and treatments received by using a digital medical carnet. Provide suggestions for possible treatments and which doctors and specialists are recommended to treat patients. This process is supported by the semantic similarity. Ontologies contain concepts that represent diseases, treatments, and the relationships among these, patients and doctors. The patient's medical card (medical history) is transferred via Bluetooth. The results were as expected according to the functionality and operation of a traditional medical card (on paper) which is used in clinics and hospitals. Also, the suggestions and recommendations of doctors and treatments facilitate decision-making to doctors and specialists in clinics and hospitals

According to the results consider the following future work:

- Extend the use of digital signatures to documents used across mobile and web applications.
- Integration of other mobile devices with Bluetooth to monitoring the condition of a patient.
- Include more concepts and relationships in the ontology to be able to pre-diagnosis
- Include a model of semantic processing in the cell phone to make suggestions of doctors and treatments through a mobile application.

References

1. <http://www.inegi.org.mx/est/contenidos/espanol/rutinas/ept.asp?t=tnf220&c=9194>
2. Paul J. Camp , James M. Hudson , Russell B. Keldorph , Scott Lewis , Elizabeth D. Mynatt, "Supporting communication and collaboration practices in safety-critical situations," *CHI '00 extended abstracts on Human factors in computing systems*, April 01-06, 2000, The Hague, The Netherlands
3. N. Bricon-Souf and C. R. Newman, "Context awareness in healthcare: A review," *Intl. journal of Medical Informatics* 76, 2--12, 2007.
4. V. Stanford, Beam me up, Dr. McCoy, *IEEE Pervasive Comput. Mag.* 2 (3) (2003) 13-18.
5. M. Munoz, M. Rodriguez, J. Favela, A. Martinez-Garcia, V. Gonzalez, Context-aware mobile communication in hospitals, *IEEE Comput.* 36 (9) (2003) 38-46.
6. S. Mitchell, M.Spiteri, J. Bates, G. Coulouris, Context aware multimedia computing in the intelligent hospital, in: *Proceedings of yhr Ninth ACM SIGOPS European Workshop*, Denmark, September 200
7. M. Munoz, V. González, M. Rodríguez and J. Favela "Supporting context-aware collaboration in a hospital: An ethnographic informed design," *lecture notes 2806*. Springer, New York, NY, 330-344
8. Stephanos Androutsellis-Theotokis , Diomidis Spinellis, "A survey of peer-to-peer content distribution technologies," *ACM Computing Surveys (CSUR)*, v.36 n.4, p.335-371, December 2004
9. H. Bludau and A. Koop, "2002. Lecture Notes in Informatics: Mobile Computing in Medicine," *Kellen Verlag*.
10. Jonas Landgren and Urban Nulden, "A study of emergency response work: patterns of mobile phone interaction," in *Proc. 2008 ACM Conference on Computer Supported Cooperative Work*
11. Kjeldskov and M. Skov., "Supporting work activities in healthcare by mobile electronic patient records," in *Proc. 2004 6th Asia-Pacific Conference on Human-Computer Interaction*
12. http://www.sciencedaily.com/videos/2006/0306-medical_records_on_your_cell_phone.htm
13. <http://www.bluecove.org/>
14. <http://www.jornada.unam.mx/ultimas/2008/11/01/diabetes-males-del-corazon-y-accidentes-causan-56-de-muertes-en-mexico-inegi>
15. <http://www.avetana-gmbh.de/avetana-gmbh/produkte/jsr82.eng.xml>

Design of a High Dynamic Range ADC by Concatenating Low Resolution Samples

Miguel Santiago Villafuerte Ramírez¹, Alfonso Gutiérrez Aldana²
and Luis Pastor Sánchez Fernández³.

Computing Research Center, National Polytechnic Institute, Mexico.

Av. Juan de Dios Bátiz s/n Col. Nueva Industrial Vallejo, C.P. 07738, México D.F.

¹Student, email: san.link@yahoo.com.mx, ²Researcher, email: lsanchez@cic.ipn.mx,

³Researcher, email: agutierr@cic.ipn.mx

Abstract. An innovative method to design a high dynamic range ADC is described. The ADC was developed to acquire audio samples for community noise monitoring with frequencies from 20Hz to 11kHz and a dynamic range which goes from 40dB_{SPL} to 140dB_{SPL}. This ADC module was created to have ADC boards which can be produced nationally with the same performance as imported boards but a lower price. By using low resolution ADCs, some analog amplification and filtering stages, and a concatenation algorithm, a good performance and cost efficient audio sampling module has been built for audio level calculations. In order to guarantee the signal readability a logarithmic legibility formula is introduced. A series of tests with different frequencies and dynamic ranges showed that the module can easily replace the current Sigma-Delta ADC boards.

Keywords: high dynamic range, analog-to-digital converter, concatenation, logarithmic legibility, audio sampling.

1 Introduction

ADCs are always evolving, specially the Sigma-Delta and the Sucesive Approximation architectures. This is mainly because the communications and sensors areas require more precise measurements with greater bandwidths. In the last few years the dynamic range has improved at 1dB per year or 1 equivalent bit every 6 years [1]. Theoric dynamic range can be described as the ratio between the highest amplitude and the smallest amplitude a signal has. Effective dynamic range is the ratio between the highest amplitude and the signal-to-noise level an instrument has.

Noise pollution is a recent problem that affects people, psychologically and physiologically, and noise monitoring is being applied at certain spots in very populated cities in order to know the noise levels and to apply measures to reduce them. The audio signals found in polluted cities have a dynamic range which goes from 20μPa to 200Pa or more, according to what a human can hear. A sound of 200Pa will damage a person's ear almost immediately. The dynamic range of these signals goes from 0dB_{SPL} to 140dB_{SPL}, when calculated with formula (1):

$$N_p = 10 \log_{10} \left(\frac{p_{rms}^2}{p_{ref}^2} \right) = 20 \log_{10} \left(\frac{p_{rms}}{p_{ref}} \right), \quad (1)$$

where: N_p is the Sound Pressure Level, p_{rms} is the current pressure level and p_{ref} is the reference pressure level which equals $20 \mu\text{Pa}$.

These audio signals are measured with an IEPE microphone which features high dynamic ranges ($140\text{dB}_{\text{SPL}}$ or more) and excellent bandwidths (20Hz to 20kHz). The electric signal of the microphone is converted to a discrete signal by using ADC modules. By designing an ADC module with high dynamic range features, further innovations can be applied to obtain a more affordable and full featured board for noise monitoring applications. The theoretic dynamic range of an ADC can be calculated with formula (2):

$$\text{ADC Dynamic Range} = 20 \log(2^n) + 1.76\text{dB}, \quad (2)$$

where: n is the number of bits of binary resolution [3].

According to the dynamic range that sound has, and using formula (2), the number of bits required from an ADC are:

$$\begin{aligned} 140\text{dB} &= 20 \log(2^n) + 1.76\text{dB} \\ n &= \frac{\left(\frac{140\text{dB} - 1.76\text{dB}}{20} \right)}{\log 2} = 22.96\text{bits} \approx 23\text{bits}. \end{aligned} \quad (3)$$

There certainly are 24 bit ADCs in the market that feature an excellent conversion performance, especially when they work with Sigma-Delta architectures [4]. The main objective of this work was to obtain a new ADC architecture which could be directly compared to existing stronger architectures and could be built by means of common ADCs and analog electronic parts, without the need of complex digital signal processes, such as averaging and approximation, or special signal processing algorithms such as the ones described in [1] and [5].

In order to determine the bandwidth and dynamic range of the ADC, some international standards such as [6] were studied and the features of some other community noise monitors were analyzed [7]. These are the main features that the ADC module had to accomplish (according to IEC 61672 Standard, class 2 of sonometers): a 20Hz to 11kHz bandwidth, a sampling frequency of 22kSps, a dynamic range of $140\text{dB}_{\text{SPL}}$, a SNR of 40dB_{SPL} , 0.3dB of precision and an analog input voltage up to $10V_{\text{rms}}$.

2 Concatenated ADC Architecture

The Concatenated ADC amplification stages are depicted in Fig. 1. The signal is amplified 16 times per stage. A low offset operational amplifier (OP177GP with $V_{OS} < 60\mu\text{V}$) was used in order to avoid the offset from being amplified stage by stage

and to prevent clipping at further stages. The fifth stage will always have a high offset, so an anti-offset filter was added to attenuate this DC value.

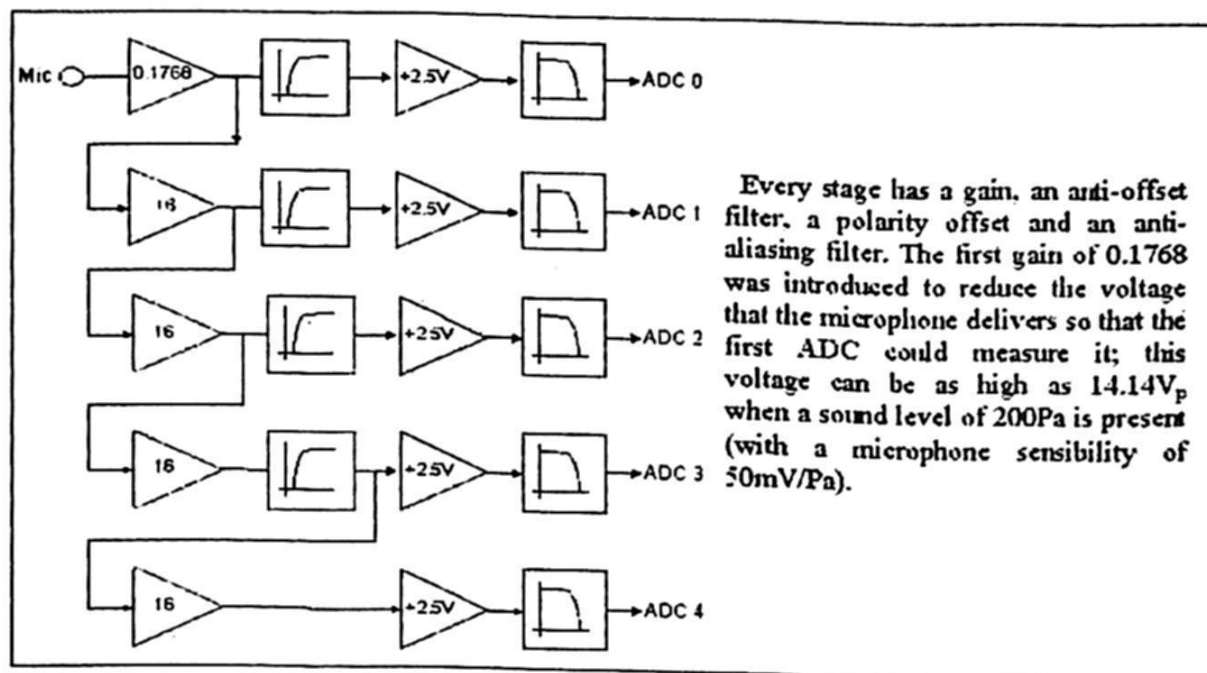


Fig. 1. Concatenated ADC amplification stages.

It is important that the signal, which is amplified in every stage, keeps a low delay between stages. If an anti-offset filter was added to the last stage, then the same filter should be added to every previous stage to keep a constant delay between them. These anti-offset filters are high-pass with a 1-order curve and a -3dB point at 20Hz.

After the signal has been amplified, a positive offset of 2.5V must be added to it so that the unipolar ADCs can measure it within a range from 0V to 5V. An anti-aliasing group of filters with a cutoff frequency at 11kHz are added just before every ADC.

If the instantaneous signal is so low to be measured by the low resolution ADC in the first stage, then the next measurement will be tested, now that the voltage is greater than the one in the first stage. If the second ADC can't measure it, then the third ADC should give it a try. And so on, until the fifth stage should be able to measure it.

In order to define when a voltage is measurable by an ADC or not, lineal and logarithmic legibility should be calculated. Lineal legibility is the smallest measurement that an ADC can present. It can be calculated by using formula (4).

$$\text{Leg}_L = \frac{V_{\text{max scale}}}{2^n - 1}, \quad (4)$$

where: $V_{\text{max scale}}$ is the highest value that the ADC can measure and n is the number of bits that the ADC has.

As a first approach, a 9 bit theoretic ADC is chosen to determine if it can present good legibilities. Its lineal legibility (also known as sensibility) is calculated with formula (4):

$$Leg_L = \frac{2.5V}{2^9 - 1} = 4.89mV / bit, \quad (5)$$

Logarithmic legibility is the smallest measurement in dB units that an ADC can present and it is calculated by using formula (6).

$$Leg_{dB} = 20 \log(V_{min\ scale} + Leg_L) - 20 \log V_{min\ scale} =$$

$$Leg_{dB} = 20 \log \left(\frac{V_{min\ scale} + Leg_L}{V_{min\ scale}} \right), \quad (6)$$

where: $V_{min\ scale}$ is the smallest value to measure on the ADC lineal scale.

If $V_{max\ scale}$ is 2.5V and a $V_{min\ scale}$ value of $(1/16)(V_{max\ scale})$ is chosen, then the logarithmic legibility is as follows:

$$Leg_{dB} = 20 \log \left(\frac{156.25mV + \frac{2.5V}{511}}{156.25mV} \right) = 0.2677dB, \quad (7)$$

where 0.2677dB are smaller than the $\pm 0.3dB$ proposed for the logarithmic precision of the ADC module. It is well known that most ADCs have noise in their 2 LSB (differential and nonlinearity issues) and the logarithmic legibility can get worsened by that. By adding 2 bits of resolution to the ADC this problem can be avoided and the Leg_{dB} can be improved:

$$Leg_{dB} = 20 \log \left(\frac{156.25mV + \frac{2.5V}{2047}}{156.25mV} \right) = 0.0676dB. \quad (8)$$

A resolution of 11 bits should be enough to guarantee a $0.3dB_{SPL}$ precision on a single polarity signal, but the microphone delivers positive and negative signals. To add negative polarity support, an extra 1 MSB should be added to the ADC, requiring then 12 bits of resolution. The ADCs chosen for this project were the MCP3201 produced by Microchip. The MCP3201 features an SAR architecture with 12 bits of resolution and 100kSps of sampling frequency.

Now that Leg_{dB} is guaranteed, Leg_L must be improved. Previous calculations showed that a 9 bit ADC has a 4.89mV/bit legibility which is not enough to measure the 100 μ V of a 40dB $_{SPL}$ signal (with a microphone sensibility of 50mV/Pa and 2mPa of sound pressure). In order to improve the Leg_L the amplification stages are used.

If an instantaneous voltage level in the first amplification stage is greater than 1/16, then that level will be legible and the ADC can measure it with good Leg_{dB} , but if the voltage is lower than 1/16, then the next ADC will have to measure it. Every stage has the same legibility set point at 1/16 as shown in Fig. 2. When the signal is taken from the last stage, it is because its magnitude was lower than 3.05mPa. A level of 40dB $_{SPL}$ or 2mPa is lower than the maximum scale at the last stage, therefore, a theoric dynamic range of 40dB $_{SPL}$ is guaranteed.

A simultaneous measurement has to be taken from every stage by every ADC. By having 5 simultaneous samples a legible sample can be chosen, so that the instantaneous signal doesn't change while the microcontroller is choosing an ADC channel.

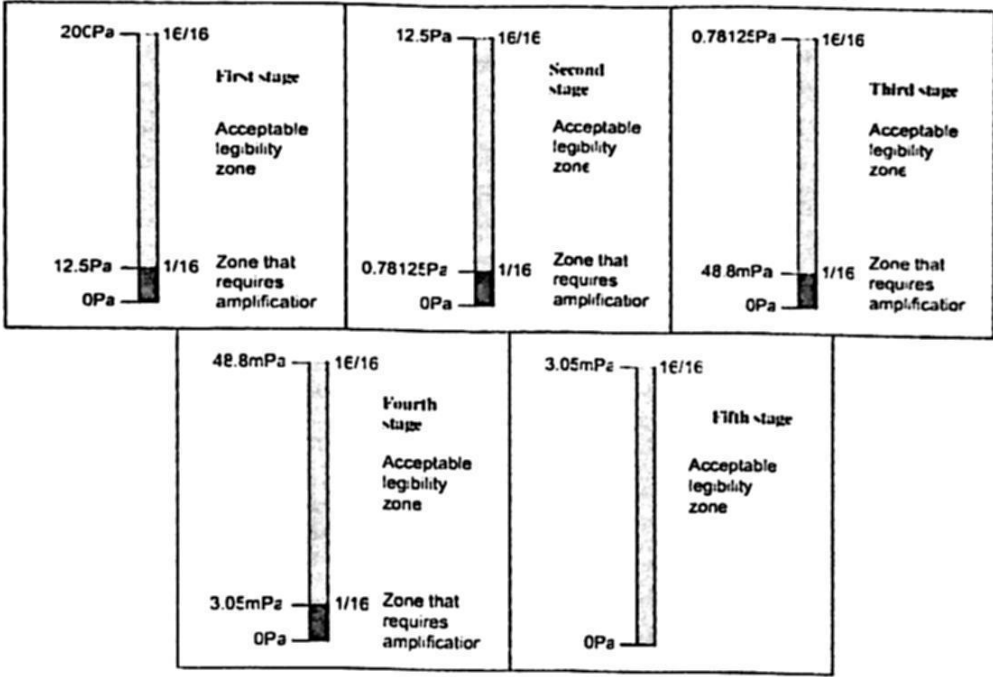


Fig. 2. Amplification stages used in a Concatenated ADC.

Choosing a legible sample from all 5 channels is easily done. Considering that every ADC has 2048 steps due to its binary resolution on one polarity, the non legible zone is delimited by 1/16 of 2048. The ADC originally has 4096 steps, where the first 2048 contain the negative polarity signal, and the last 2048 ones contain the positive one (see Fig. 3).

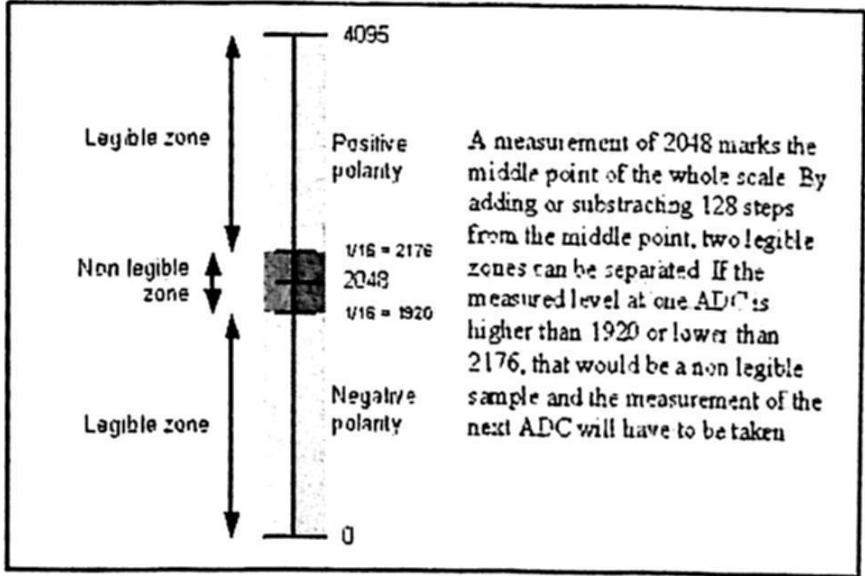


Fig. 3. Legible zones in both polarities.

Once that a legible sample has been found it can be concatenated into a higher resolution sample by adding zeroes depending on which stage it was located and taking advantage of the 2^4 gain. For instance, if a sample had a level of 2mPa it would be located in the fifth stage as seen in Fig. 4 (a). The microcontroller will take the measurement of the first stage and will determine if that level is lower than 1/16 of the V_{max} scale. It is known that every bit in a binary sample equals one half of a scale. If the MSB is 1, the sample is located at the upper half, and if it's 0 it's located in the lower half. The current measurement is lower than 1/16, and then the MSB will be 0. The next bit represents a fourth of the scale, it will be a zero since the sample is still lower than 1/4. The next two bits, the eighth and sixteenth portions will also be zeroes. Now it is known that the first ADC does not have enough Leg_{dB} to measure the sample.

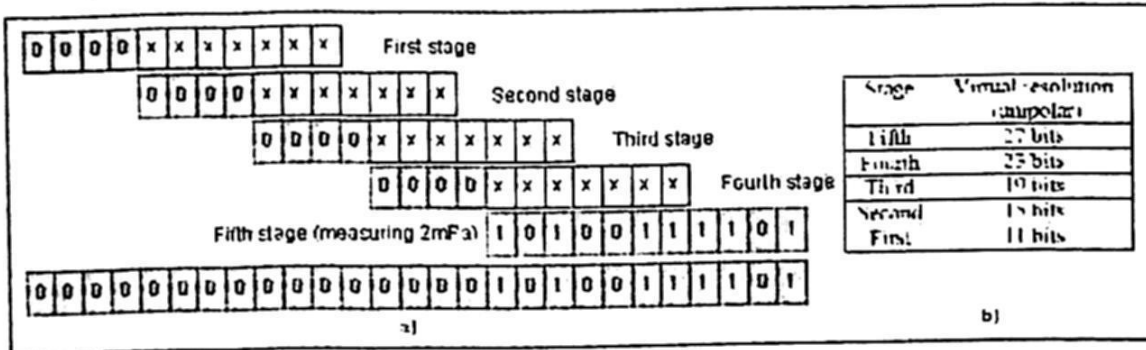


Fig. 4. Concatenation of a legible sample into non legible samples.

From stage 1 to stage 4, there will be 4 zeroes from every ADC and the last stage will have 11 legible bits. According to these facts, a virtual improved resolution is obtained depending on which stage contains the legible sample. The stage resolution is showed in Fig. 4 (b). An extra MSB must be added to indicate the signal polarity; therefore, a maximum of 28 bit virtual resolution can be obtained if a signal is located at the fifth stage. The virtual lineal legibility is calculated as follows:

$$Leg_L = \frac{(10V_{rms})(\sqrt{2})}{2^{27} - 1} = 105.36nV/bit, \quad (9)$$

which is lower than the $100\mu V$ that a $40dB_{SPL}$ signal has.

The 12 bit resolution and the minimum scale voltage of 1/16 are a good combination since 12 bit ADCs and an array of 5 amplification stages are easy to implement. The greater the number of amplification stages implemented, the lower ADC resolution is required. The greater the ADC resolution used, the smaller the logarithmic legibility.

3 ADC Module Performance

Once that the ADC module was assembled, a series of dynamic range and bandwidth tests were performed with a signal generator. An application was programmed with Visual C# 2008 in order to connect the board via USB, stream the legible samples,

calculate the equivalent continuous sound level and plot the signals in time and frequency domains. The frequency domain plot was calculated with regular windows such as Hamming and Hanning, or the Nuttall high dynamic range window [8].

Table 1 shows the frequency response of the module when tested with a 114dB_{SPL} signal. The -3dB frequency begins at 20Hz due to the anti-offset filter and the second -3dB point is found at 11kHz according to the anti-aliasing filter (see Fig. 5). A 20Hz to 11kHz bandwidth was achieved with an acceptable precision.

Table 1. ADC module frequency response.

Frequency (Hz)	dB _{SPL}	Gain (dB)	Frequency (Hz)	dB _{SPL}	Gain (dB)
10	107.17	-6.83	800	113.99	-0.01
20	111	-3	1000	113.99	-0.01
30	112.39	-1.61	1250	113.97	-0.03
50	113.32	-0.68	1600	113.93	-0.07
80	113.73	-0.27	2000	113.91	-0.09
100	113.84	-0.16	2500	113.82	-0.18
150	113.95	-0.05	3150	113.71	-0.29
160	113.96	-0.04	4000	113.5	-0.5
200	113.98	-0.02	5000	113.29	-0.71
250	113.99	-0.01	6300	112.97	-1.03
315	114	0	8000	112.39	-1.61
400	114	0	10000	111.5	-2.5
500	113.99	-0.01	10900	111.07	-2.93
630	114	0			

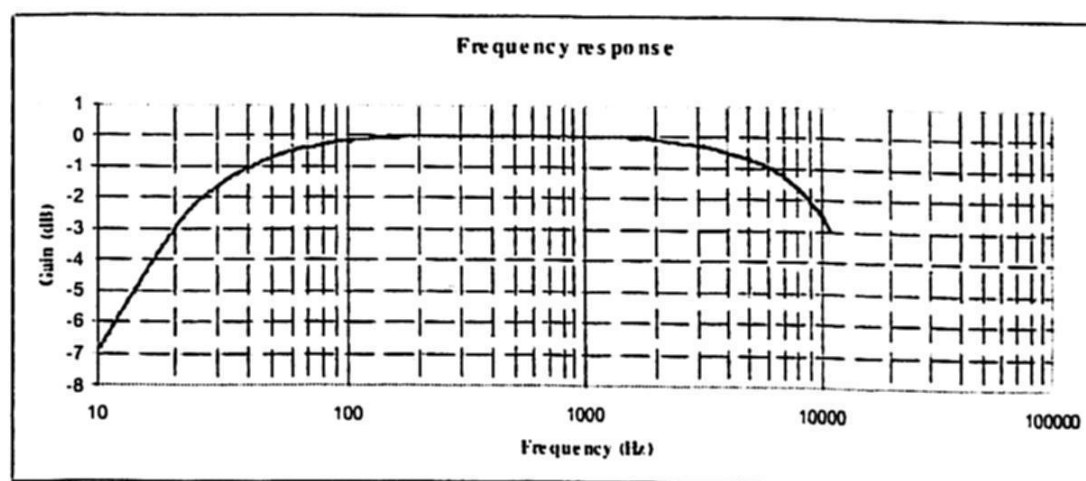


Fig. 5. Frequency response of the concatenated ADC module.

The signal-to-noise level of the module was tested with a 0V signal. An image of this signal is shown in Fig. 6. A level of 35dB_{SPL} was found, which is better than the 40dB_{SPL} proposed.

Further levels were tested with a constant frequency signal of 1kHz and a dynamic range from 49dB_{SPL} to 135dB_{SPL} (see Fig. 7 and 8). Distortion levels were found in the range from 49dB_{SPL} to 69dB_{SPL} due to the very small amplitude that the signal generator had to output, but the signals were correctly measured and plotted.

Some phase delays were observed between stages, when a stage was switched to a higher or lower one. These delays were given by the difference between the RC values on each anti-offset filter and by the opamp delay. Although the phase delay of the original signal is affected by the anti-offset and anti-aliasing filters, there are no definitive studies that demonstrate that phase delays significantly affect the quality of the signal when heard by a human [9], so the digital signal can still be used to measure noise levels.

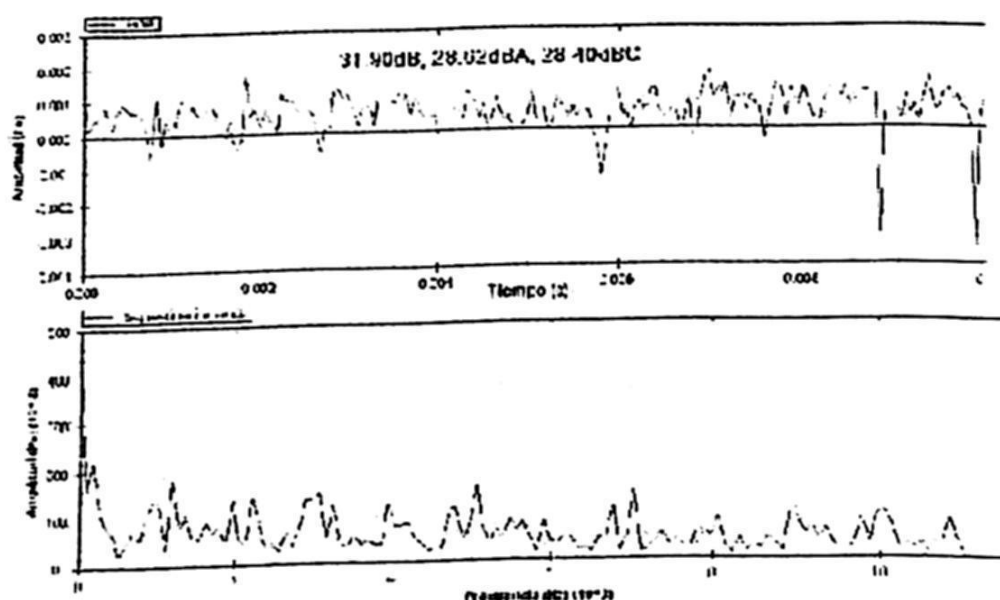


Fig. 6. Signal-to-noise level of the ADC Module.

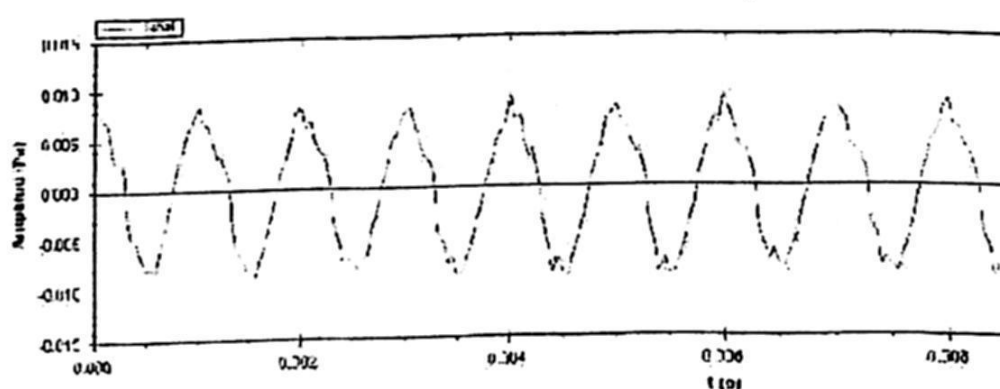


Fig. 7. A 49dB_{SPL} (500μV @ 1kHz) signal measured by the ADC module.

A theoretic dynamic range of 167dB was calculated by using formula (2):

$$\text{Dynamic Range} = 20 \log(2^{27}) + 1.76 \text{ dB} = 164.3 \text{ dB}. \quad (10)$$

An effective dynamic range of 105dB was obtained by subtracting 35dB (SNR) from 140dB. The bandwidth (20Hz to 11kHz) and dynamic range (40dB_{SPL} to 140dB_{SPL}) were accomplished with low resolution ADCs and simple analog filtering and amplification stages.

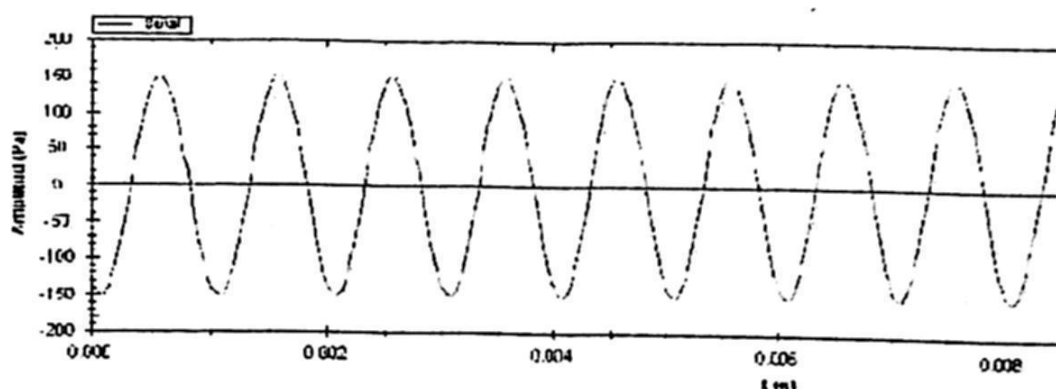


Fig. 8. A 135dB_{SPL} (6.84V @ 1kHz) signal measured by the ADC module.

4 Discussion

The applied tests on the ADC module showed that it can measure audio levels from 20Hz to 11kHz with very good precision. This ADC module can be compared to other commercial modules with similar features. Table 2 shows these features.

The National Instruments NI USB 9233 is an over-featured board when used for noise monitoring, since audio sampling requires up to 32kSps (according to [6]) and only one input channel. A sampling frequency of 22kSps is enough since some noise monitoring experiments that have been done, showed that signals with significant amplitudes are only present at up to 8kHz.

The theoretic dynamic range of the Concatenated ADC is better than the NI USB since it has more resolution bits, but the effective dynamic range is almost similar to the NI USB. Both ranges can compress an audio signal ranging from 40dB_{SPL} to 140dB_{SPL}.

Table 2. Concatenated ADC module vs. NI USB 9233.

	NI USB 9233	ADC Module Prototype
Sampled channels	4	1
Sampling frequency	up to 50kSps	22kSps
Resolution	24 bits	27 bits plus sign
Theoric dynamic range	146dB	164dB
Effective dynamic range	98dB (25kSps)	105dB (22kSps)
ADC architecture	Delta-Sigma	Concatenated ADC
AC Cutoff Frequency	0.5Hz	20Hz
Input Voltage Range	$\pm 5.8V_p$	$\pm 14.14V_p$
Precision	0.6dB (no calibration)	± 0.2677 dB, according to L_{eqdB}

The sampling frequency of the prototype was limited due to the PIC microcontroller used, which has a 12MIPS speed. If a faster microcontroller or a digital signal processor was used, then a 32kSps speed could be achieved and a Class 1 Sonometer could be integrated within the module.

5 Conclusions

The concatenated ADC that has been proposed and designed in this work can be used on signals that have bandwidths a little beyond the 0Hz frequency due to the anti-offset filtering required. No DC measurements can be done with this ADC architecture. The minimum scale voltage and the number of ADC resolution bits mark the balance between amplification stages and logarithmic legibility.

A phase delay between stages can be made almost zero if the RC values are similar between stages. This is achievable by implementing low tolerance elements with 1% of precision or less.

A virtual resolution of 27 bits plus sign was obtained by concatenating the measurements of each stage. Simultaneous ADC samplings are required in order to prevent samples loss and effective concatenations.

These concatenations imply the adding of zeroes to the left or right of the legible sample if the minimum scale voltage is a power of 2 (for example, 1/16 or 1/32). A microcontroller can easily add these zeroes by shifting the value to the left or right. When a sample is found legible in the first stage, zeroes are added to the right. When it is found on the second, third or fourth stages, zeroes are added on both sides. When it is found on the last stage, zeroes are added to the left.

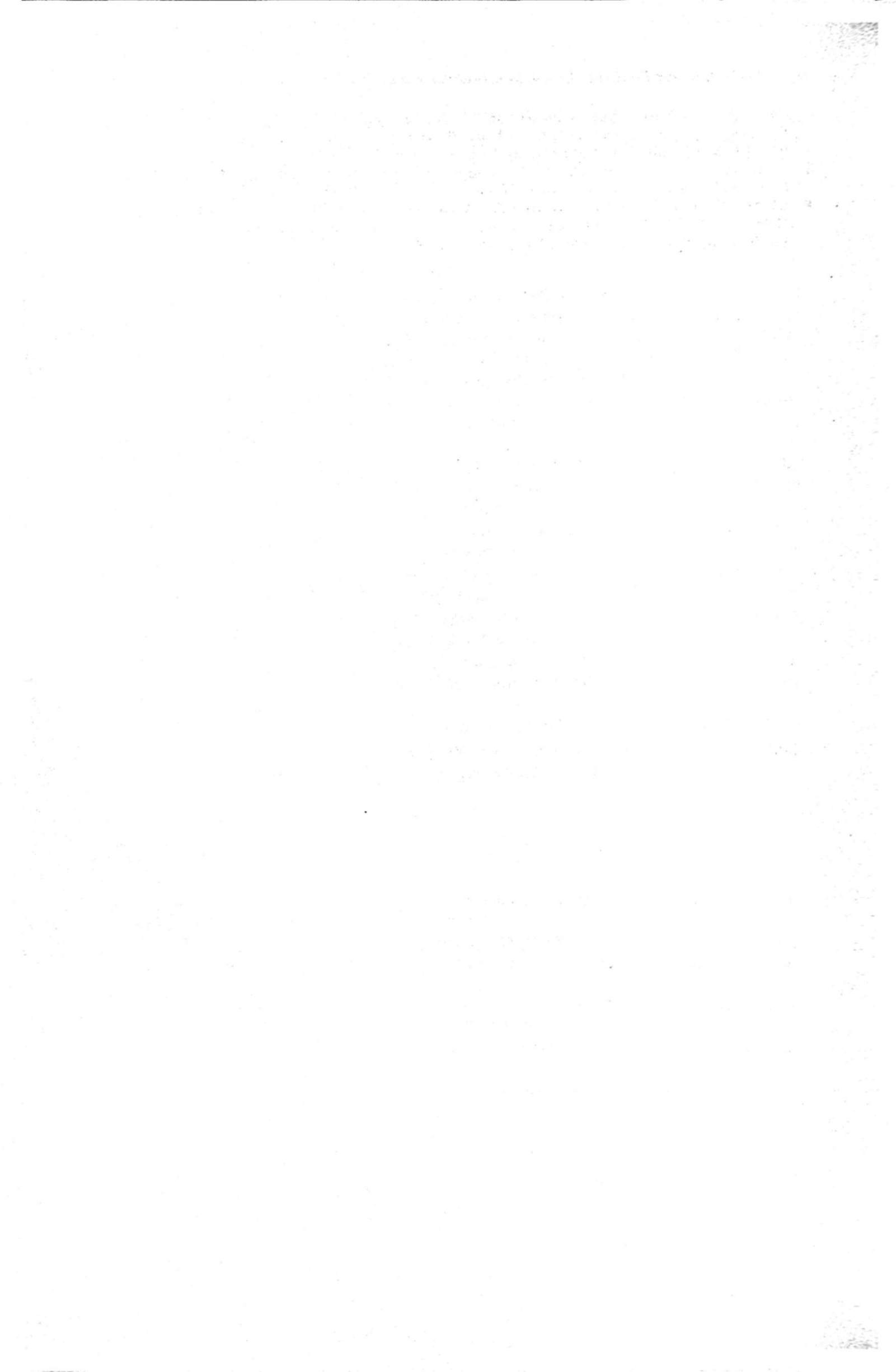
Logarithmic legibility indicates whether a signal can be correctly measured in a logarithmic scale and achieve a good precision, depending on the application requirements.

A Bessel analog filter can be implemented instead of the 1-order anti-offset filter used. This filter would have a phase delay present beyond the bandwidth of the signal to be measured. The phase delay between stages would almost disappear.

References

1. SONG, W. S. High dynamic range Analog-To-Digital Converter having parallel equalizers. United States Patent US 6,653,959 B1, p. 7, (2003)
2. INTERNATIONAL ELECTROTECHNICAL COMMISSION. *Norma IEC 61094-4 Micrófonos de medida*. Parte 4 - Especificaciones para los micrófonos patrones de trabajo, 1a ed., (1995)
3. ROHDE&SCHWARZ. dB or not dB? Everything you ever wanted to know about decibels but were afraid to ask (Application Note IMA98) (2006)
4. SCHREIBER, R. Precision Analog-To-Digital Converter Architecture Trade-Offs. ChipCenter. Texas Instruments.
5. FETTE, B. A. High Dynamic Range Analog To Digital Converter Architecture. United States Patent US 7,253,755 B1, p. 4, (2007)

6. INTERNATIONAL ELECTROTECHNICAL COMMISSION, *Norma IEC 61672-1 Electroacústica Sonómetros. Parte 1 - Especificaciones*, 1a. ed., p. 44, (2005)
7. BRUEL & KJAER. Noise Monitoring Terminal Type 3597 (2004)
8. GUMAS, C. C. Window choices become crucial in high-dynamic-range FFT processing. *Personal Engineering & Instrumentation News*, p. 53-59 (1997)
8. SHANEFIELD, DANIEL; LIPSHITZ, STANLEY P.; POCOCK, M.; VANDERKOOY, JOHN. On the Audibility of Midrange Phase Distortion in Audio Systems. *Audio Engineering Society*, Princeton, NJ, p. 447-448 (1983)



A Software Tool for the Analysis of Similarity in Recurrence Patterns

Ernesto Bautista-Thompson¹, Roberto Brito-Guevara¹,
Jesús E. Molinar-Solis²

¹Centro de Tecnologías de la Información, DES-DACI, Universidad Autónoma del Carmen, Calle 56 Número 4, C. P. 24180, Ciudad del Carmen, Campeche, México
²Centro Universitario UAEM-Ecatepec, Universidad Autónoma del Estado de México
José Revueltas 17, Col. Tierra Blanca, C. P. 55020, Ecatepec de Morelos,
Estado de México, México
teb_thompson@yahoo.com, rbritoguevara@hotmail.com,
molinarov@hotmail.com

Abstract. Recurrence plots visualize spatial and also in the case of time series temporal correlations inside sequences of data, the technique allows the identification of hidden data relationships (periodicity, non-stationarity, recurrence, randomness) inside a sequence. The comparison between recurrence patterns in order to identify common structures is difficult because the lack of similarity quantification tools in the available and most popular software for recurrence plots analysis such as VRA (Visual Recurrence Analysis), RQA (Recurrence Quantification Analysis) and CRP (Cross Recurrence Plots). In this work a software tool for analysis of structural similarity patterns between recurrence plots is proposed, this tool named RecurrenceVs, allows the comparison and quantification of the degree of structural similarity between recurrence plots generated from different sequences of data. The results shows that this tool is useful for the classification of data sequences by similarity families based on the recurrence patterns, where these patterns preserve the information about the structure and dynamics of data sequences.

Keywords: Recurrence Patterns, Spatial and Temporal Correlation, Structural Similarity.

1 Introduction

The extraction of common structural features in sets of time series and sequences of data is an important task in the identification of patterns of interest for example: in the analysis of time series dynamics [1], the identification of motifs in genomics sequences [2], the construction of queries for sequence extraction in databases [3]. Similarity in data sequences can be measure with different metrics such as Euclidean distance [4], Dynamic Time Warping [5], Similarity Histograms [6], etc. such measures operates directly over the data sequences. For other side, analysis of the dynamics of sequences of data, understood as the relationships between the different data inside a sequence, can be done with different techniques such as: Autocorrelation [7], Wavelet Analysis [8], Recurrence Quantification Analysis [9], etc. In particular, the

Recurrence Quantification Analysis is based on the generation of a data representation named recurrence plots, these plots shows patterns of hidden relationships between data such as non-stationary behavior, periodicities, and randomness; the patterns generated with these plots can be used to compare sequences of data at a new level of information. There are not quantitative tools for comparison of recurrence patterns, software such as VRA, RQA and CRP lacks this functionality [10, 11, 12], this is the main motivation for the development and proposal in this work of a software tool for quantitative comparison and analysis of similarity between recurrence patterns. Section two, explains the theoretical basis of the recurrence plots as well as some concepts of similarity. Section three, describes the technical aspects of the RecurrenceVs software tool. In Section fourth, the experimental results of the evaluation of the software tool are presented. Finally in section five, the discussion of this work is presented.

2 Recurrence Plots

Recurrence plots are graphical representation of a sequence of data, which allows the detection of hidden dynamical patterns and nonlinearities inside the data. These plots allow the visualization of recurrent patterns, non-stationary patterns, and structural changes. The recurrence plot is part of a technique known as Recurrence Analysis that was developed by Eckman et al. in 1987 [9].

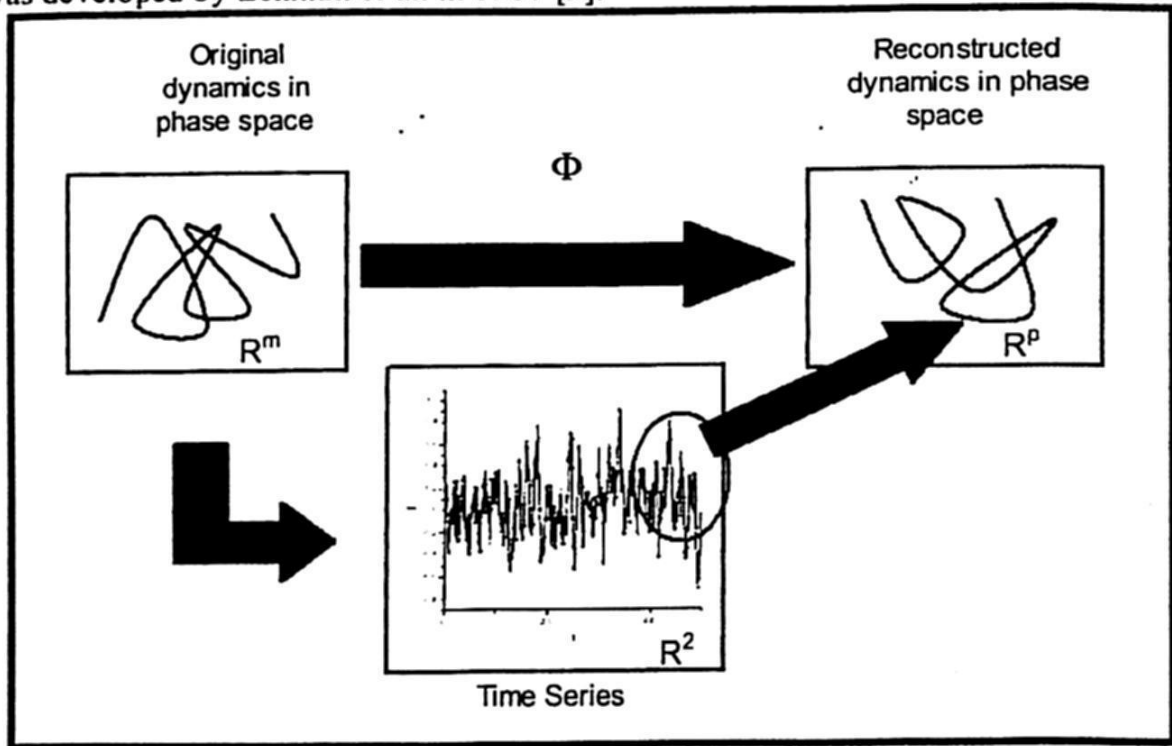


Fig. 1. Conceptual schema of reconstruction of a multidimensional system based on an observable (time series).

The theoretical foundation of the construction of recurrence plots is based on two theorems proposed by H. Whitney and F. Takens [13, 14], in these foundational works they establish that is possible to recreate a topological equivalent of the original

behavior from a multidimensional system by means of a sequence of data from an observable of such system [14], the Fig.1 illustrates this idea.

In the general case of the reconstruction of a system with embedding dimension n and a delay $\tau=0$, each datum $x(i)$ is a vector composed by n consecutive data elements from the sequence.

$$x(i) = \{x(1), x(2), x(3), \dots, x(n)\}. \quad (1)$$

A recurrence plot is generated by comparison of each datum in a sequence with itself and with the rest of data. The comparison between, for example, data $x(i)$ and $x(j)$ is made with a metric such as Euclidean distance.

$$d_{ij} = \|x(i) - x(j)\|. \quad (2)$$

This allows building a correlation matrix D of spatial and temporal nature (as is the case of time series).

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{bmatrix}. \quad (3)$$

Each element d_{ij} in matrix D is associated with the Euclidean distance between a datum in position i and a datum in position j inside a sequence, if a datum or subsets of data are recurrent this behavior will be detected by means of sets of equal distances. Also, each distance d_{ij} is associated with a value from a discretized gray tone scale, in this way an image of the recurrence plot pattern is generated. The resultant pattern is symmetric due to the redundancy in the calculation of the distances $d_{ij} = d_{ji}$.

Comparing the patterns generated for the recurrence plots can be useful in order to identify sequences of data with similar data behavior or dynamics, this similarity must be quantified in order to have an objective comparison of the recurrence plots.

3 Design of RecurrenceVs: Architecture and Algorithms

The design of the software tool has two important aspects: user interface design and algorithm design. In the first case, the user interface must be easy of use and to be windows driven, the software must allow the execution of a series of multiple experiments and it must be easy to save the results in incremental way. The user interface features were motivated by the analysis of different software tools for recurrence plot analysis (VRA, RQA, CRP), in Table 1, a comparative feature analysis between these

software tools is showed and in Fig. 2 a screenshot of the RecurrenceVs interface is presented.

Table 1. Comparison of features between different tools.

Comparison of Features for Recurrence Analysis Software			
Software	User Interface	Similarity Analysis Function	Multiple Experiments in one Run
RQA	DOS Command Line	No	Yes, in Batch Mode
VRA	GUI	No	No
CRP	MatLab Command Line	No	Needs Programming
RecurrenceVs	GUI	Yes	Yes, in User Sesion

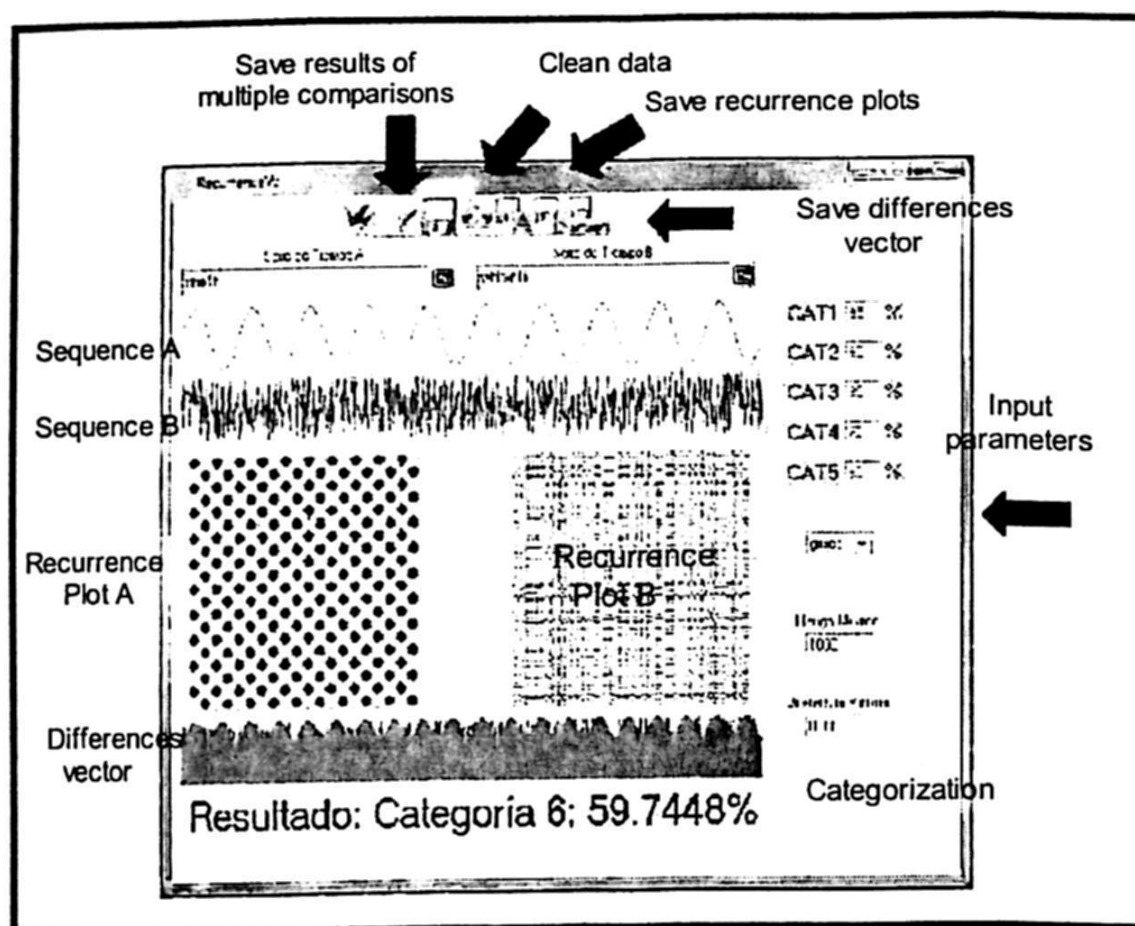


Fig. 2. Screenshot of the software interface and its elements.

The design of the algorithms was divided in three stages: first the load and preparation of the data, second the construction and comparison of the recurrence plots and third the visualization and storage of the results. In the first stage, an important aspect if the normalization of the data sequences to be compared, the data sequences can

have different length, and data range (the data types used are: integer and real), in order to generate comparable recurrence plots they are normalized to a default length of 1000 data by sampling each sequence, but this parameter of length can be adjusted if shorter time series are used in the similarity analysis. The normalization in range is done by scaling the values in order to have a range of values between 0 to 1, the Fig. 3 shows a diagram of this stage.

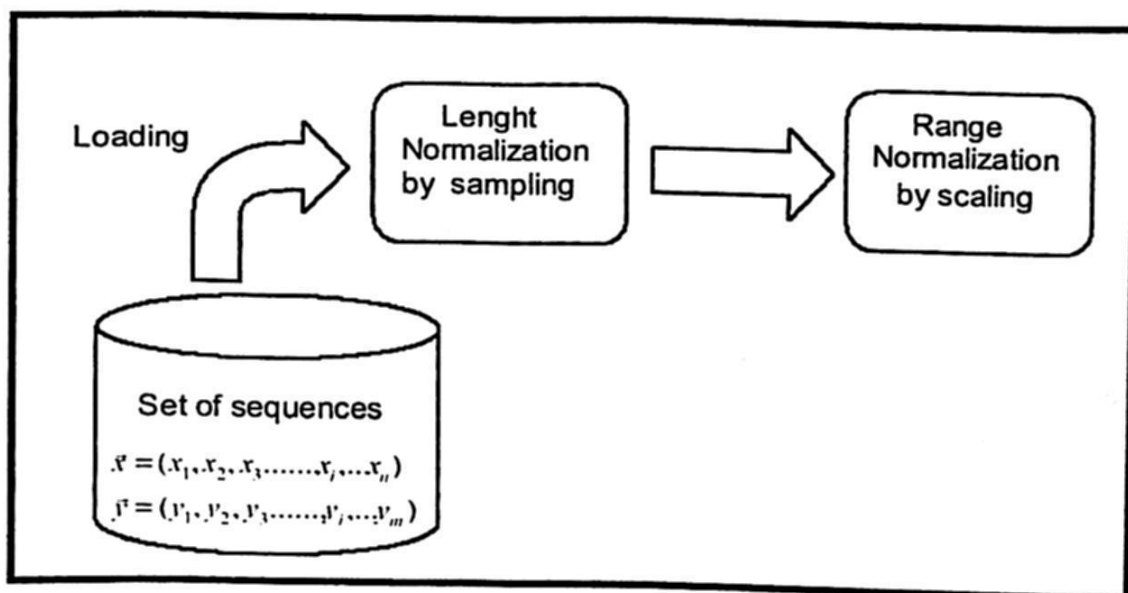


Fig. 3. Stage one, preprocessing.

The construction of the recurrence plots is done with the algorithm described in Section two, the comparison between the recurrence plots is done by comparing the corresponding element $[i, j, d_{ij}]$ of the matrix of distances for each of the two recurrence plots, then for each position (i, j) the difference between the corresponding distances d_{ij} for both recurrence plots is calculated, a difference of distances $D_{d_{ij}}$ is considered as a match between both recurrence plots if it has a value below a similarity threshold S_r , this input parameter is setup by the user before the beginning of the analysis, the threshold allows to establish different degrees of similarity between the recurrence plots, the values of the differences between the distances d_{ij} are discretized as 1 for those below the threshold and 0 above the threshold and storage in a distance vector, a counting of the number of minimum differences is done, in order to determine the percentage of minimum distances between the compared recurrence plots, a set of five similarity percentages are established by the user, and depending on the corresponding percentage reached by the counting process a similarity degree between the recurrence plots is determined. The Fig. 4 shows the steps of this second stage.

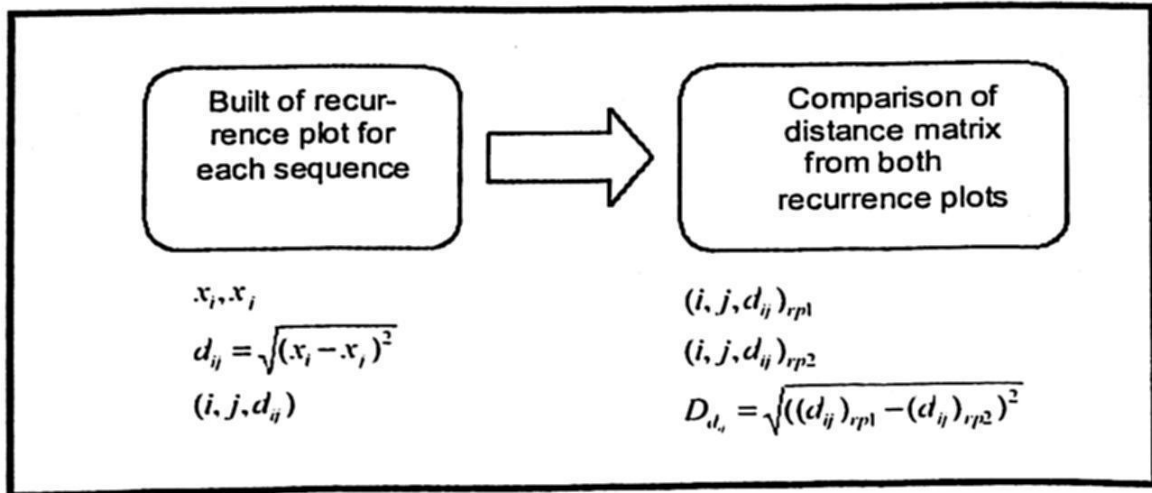


Fig. 4. Stage two, recurrence plot comparison.

Finally, in the third stage the visualization of the recurrence plots and their results are showed, the recurrence plots are visualized using a gray scale that corresponds with the distance between their points, the white color represents the minimum distance and the black color corresponds to the maximum distance present in the plot, each graphic of a recurrence plots is generated with a sampling of their corresponding distance set in order to facilitate its visualization, this is due to the enormous quantity of data generated, for example for a sequence of 1000 data a matrix of 10^6 points is generated. In this stage, the option to save the recurrence plots and the results of their comparison is activated, multiple comparison experiments can be done in one run for example: comparing a specific recurrence plot against a set of different recurrence plots can be saved by incremental storage of each new experiment, in the Fig. 5 is showed the steps of this last stage.

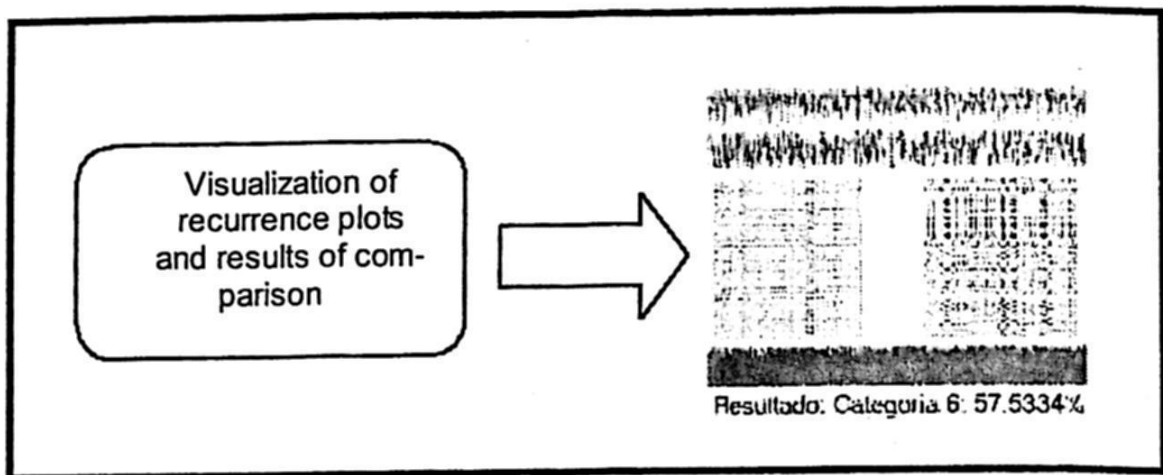


Fig. 5. Stage three, visualization of the results.

4 Evaluation: Performance and Usability

The software tool was evaluated with a set of synthetic data sequences previously classified by their similarity degree by means of the direct comparison of the sequences using the technique of Derivative Dynamic Time Warping (DDTW) [15, 16]. Examples of different comparisons are showed in the screenshots of Fig. 6 and Fig. 7; in these examples a recurrence plot from a sequence tagged DS11 is compared with the corresponding recurrence plots for the sequences DS12 and DS42; in the previous classification DDTW reported in [15, 16], the similarity of these sequences groups DS11 and DS12 in the same class 5 and the sequence DS42 belongs to class 1. The results with RecurrenceVs show a correspondence with the aforementioned results.

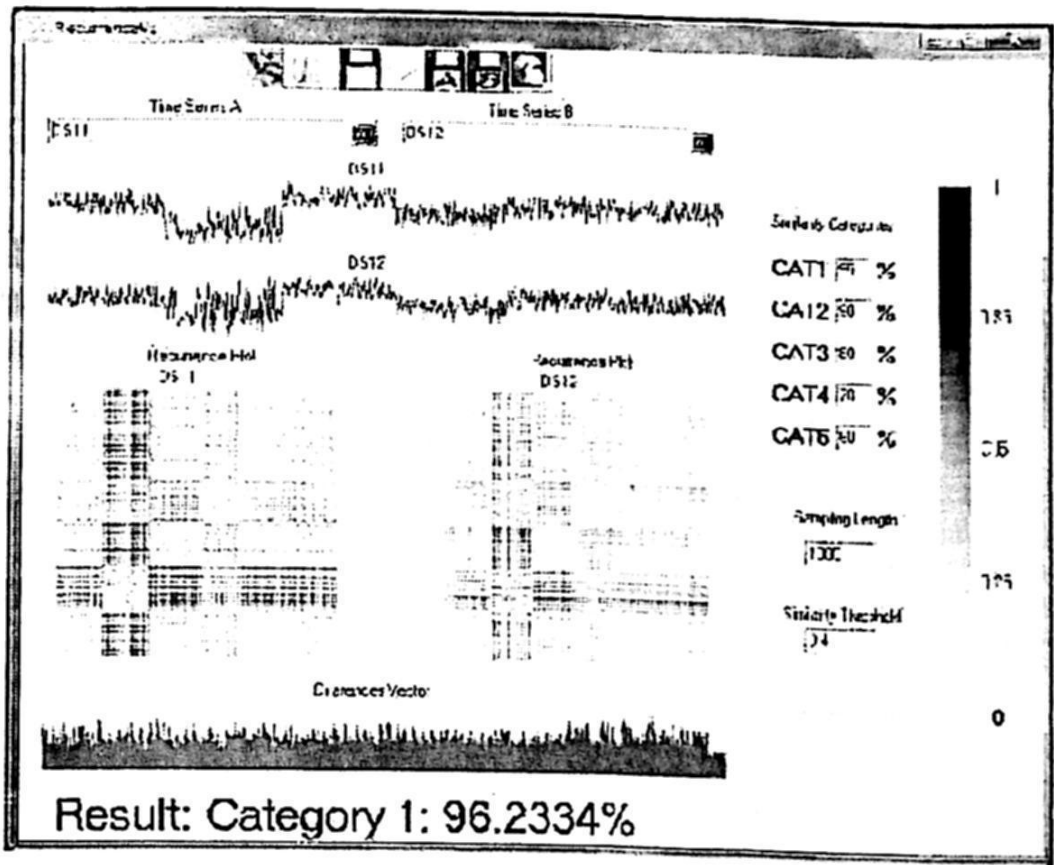


Fig. 6. Screenshot of the similarity analysis for the recurrence plots of two synthetic data sequences DS11 and DS12, their similarity corresponds to category 1 (95% of similarity).

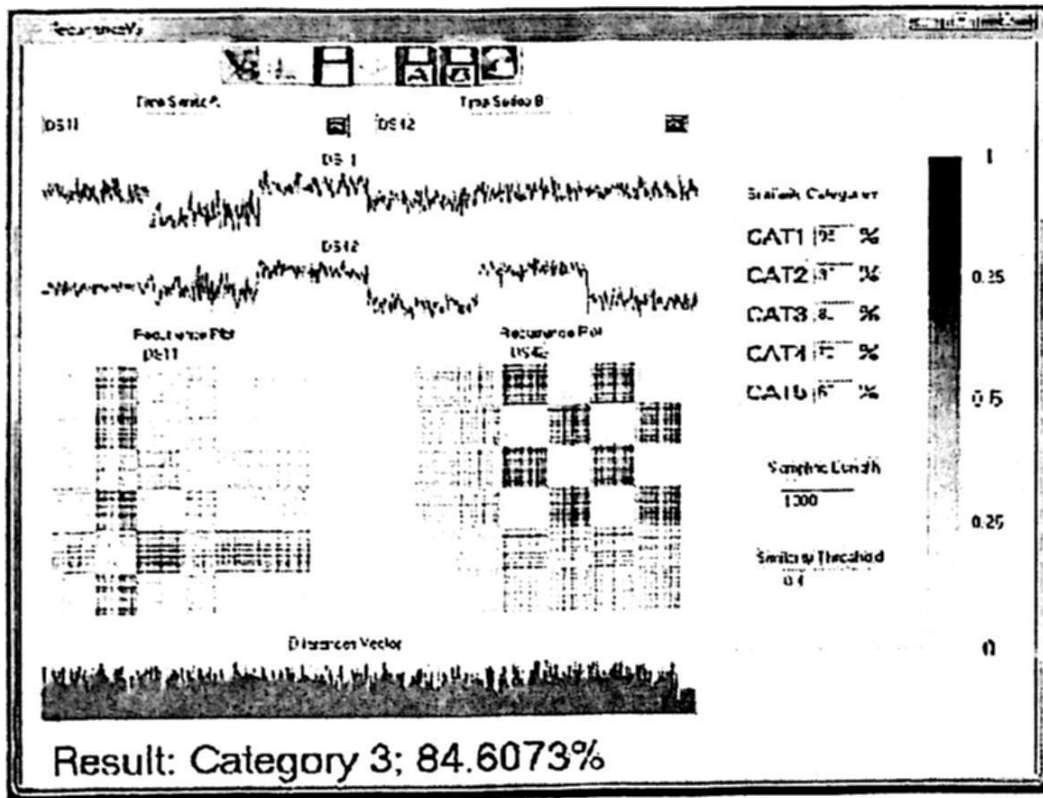


Fig. 7. Screenshot of the similarity analysis for the recurrence plots of two synthetic data sequences DS11 and DS42, their similarity corresponds to category 3 (84.6% of similarity).

In Fig. 8 is show, the behavior of the similarity percentage for the similarity analysis made with the RecurrenceVs tool, for the sequence DS12 that belongs to the class 5 in the DDTW classification [15, 16].

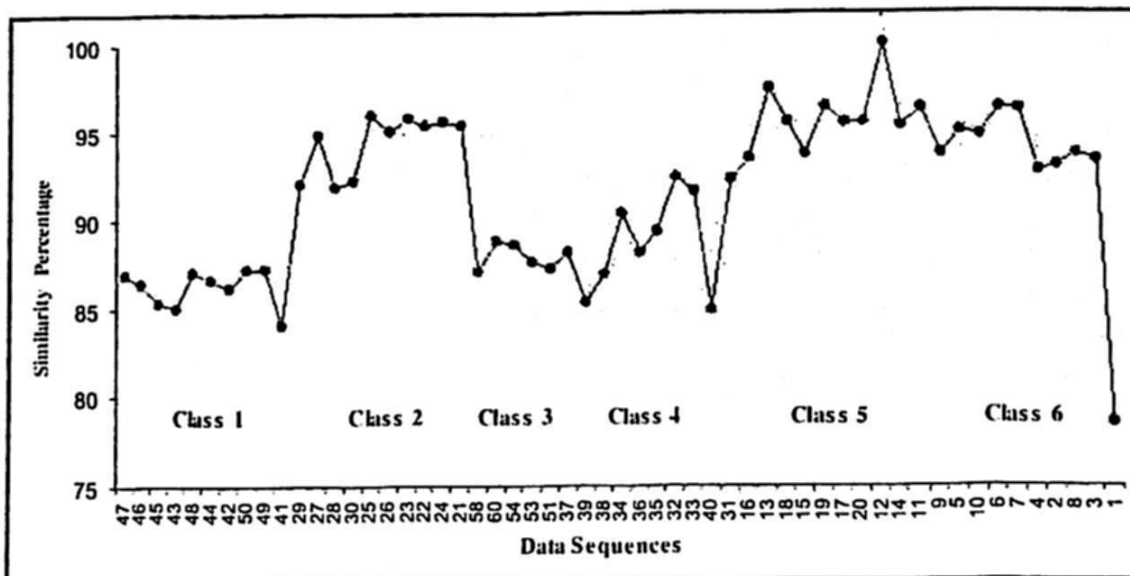


Fig. 8. Behavior of the similarity percentage between the recurrence plots for the data sequence DS12 and the rest of the sequences from the experimental set.

The sensibility of the input parameters is illustrated in Table 2, it shows the effect that changes in the similarity threshold S , have on the categorization of the recurrence

plots, in this example for the sequences DS11 and DS12, the sensibility is greater for a value of $S_r = 0.1$, this corresponds with a requirement of almost identical recurrent plots, for values of S_r between 0.3 and 0.4 the sensibility is reduced and no change in the categorization is observed.

Table 2. Sensibility of input parameters, categorization of the compared recurrence plots for the DS11 and DS12 similar data sequences.

Sensibility of input parameters			
Comparison of two similar sequences			
Similarity threshold	Assigned category	Percentage of similarity	Threshold of similarity percentage
0.1	6	57.28%	below 60%
0.2	3	86.55%	80%
0.3	1	97.17%	95%
0.4	1	99.59%	95%
0.1	5	57.28%	50%
0.2	2	86.55%	80%
0.3	1	97.17%	90%
0.4	1	99.59%	90%

5 Discussion

A software tool for the analysis of similarity between recurrence patterns was developed, this tool identified as RecurrenceVs includes a user-friendly interface in order to develop a series of experiments for the study of similarity between sets of recurrence plots where these represents different dynamical behaviors from sequences of data. The evaluation of the software tool with the set of time series from [15, 16] shows that it is capable of discriminate between similar and non similar sequences based on their recurrence representations and generate classifications based on the recurrence patterns. The parameterization of the similarity by means of a threshold and a similarity percentage is useful because it allows the analysis of similarity between recurrence patterns where their differences are not sharp. This analysis tool can be a complement to the existing recurrence analysis tools (VRA, RQA, CRP) where different properties such as percent of recurrence, percent of determinism, Shannon entropy, etc. can be calculated for each recurrence plot and in this way correlate the similar recurrence patterns with such properties.

References

1. Bautista-Thompson, E.: Measurement of Time Series Predictability: An Experimental Study. Ph. D. Thesis. CIC-IPN, México D.F. (2005).
2. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147, 195--197 (1981)
3. Das, G., Gunopulos, D., Mannila H.: Finding Similar Time Series. In: Komorowski, H.J., Zytkow, J.M. (eds.) PKDD'97, LNCS, vol. 1263, pp. 88-100. Springer-Verlag, London (1997).
4. Yi, B.K., Faloutsos, C.: Fast Time Sequence Indexing for Arbitrary Lp Norms. In: Proceedings of the 26th International Conference on Very large Databases, pp. 385-394. Morgan Kaufmann, San Francisco (2000).
5. Keogh, E., Pazzani, M.: Scaling Up Dynamic Time Warping for Data Mining Applications. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 285-289. ACM Press, Boston (2000).
6. Bautista-Thompson, E., Santos-De la Cruz, S.: Shape Similarity Index for Time Series Based on Features of Euclidean Distances Histograms. In: Gelbuck, A., Suárez-Guerra, S. (eds.) Proceedings of the 15th International Conference on Computing, pp. 60-64. IEEE Computer Society Press, Los Alamitos (2006).
7. Proakis, J. G., Manolakis, D. K.: Digital Signal Processing Principles, Algorithms, and Applications. Prentice Hall (2006)
8. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, Burlington (2009)
9. Eckmann, J.P., Kampshort, S.O., Ruelle, D.: Recurrence Plots of Dynamical Systems. *Europhys. Lett.* 4, 973-977 (1987).
10. Visual Recurrence Analysis Software (by Eugene Kononov), <http://www.myjavaserver.com/~nonlinear/vra/download.html>
11. Zbilut, J.P., Weber, C.L.: Embeddings and Delays as Derived from Quantification of Recurrence Plots. *Phys. Lett. A* 171, 199-203 (1992).
12. Marwan, N.: Encounters with Neighbours: Current Developments of Concepts Based on Recurrence Plots and Their Applications, Ph. D. Thesis. Potsdam University, Potsdam (2003).
13. Whitney, H.: Differentiable Manifolds. *Annals of Mathematics* 37, 645-680 (1934).
14. Takens, F.: Detecting Strange Attractors in Turbulence. *Lecture Notes in Mathematics* 898, 366-381 (1981).
15. Keogh, E., Pazzani, M.: Scaling Up Dynamic Time Warping for Data Mining Applications. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 285-289. ACM Press, Boston (2000).
16. Hettich, S., Bay, S. D.: The UCI KDD Archive, <http://kdd.ics.uci.edu>.

Intelligent and Adaptive User Interfaces for Ubiquitous Learning

Héctor Antonio Villa Martínez, Francisco Javier Tapia Moreno

Departamento de Matemáticas
Universidad de Sonora
Rosales y Transversal S/N
Hermosillo, Sonora 83000
{hvilla,ftapia}@gauss.mat.uson.mx
<http://www.mat.uson.mx/>

Abstract. Advances in mobile computing and in communications networks have allowed the possibility of u-learning (ubiquitous learning), which is the union of two educational systems: adaptive e-learning (electronic learning) and m-learning (mobile learning). One way of implementing a u-learning platform is utilizing an intelligent and adaptive user interface that adapts itself as the student's knowledge evolves, while helping him/her with tasks like intelligent searches and recommendations of learning material. The implementation of these user interfaces raise problems, like building a model of the student, which we propose can be solved using Bayesian networks. The purpose of this paper is twofold. First, to present our research in this area. Second, to report our first steps in constructing a u-learning platform for learning statistics.

Key words: Adaptive User Interfaces, Intelligent User Interfaces, Ubiquitous Learning, Mobile Learning, E-learning, Bayesian Networks.

1 Introduction

U-learning (ubiquitous learning) is a learning system that allows a student to receive personalized instruction anywhere and anytime. Thus defined, u-learning can be seen as the union of two learning systems: adaptive e-learning (electronic learning) and m-learning (mobile learning). Adaptive e-learning allows students to receive personalized instruction on a desktop computer by means of Internet, an intranet, a CD-ROM, or any other tool. On the other hand, m-learning (mobile learning) allows students to receive instruction anywhere and anytime on a portable computer by means of communication technologies like Wi-Fi or GSM.

A platform for u-learning can be implemented by means of an adaptive user interface that adapts itself to the students level of knowledge, computer skills, preferences, and to the features of the computer being utilized in the learning process. Furthermore, the interface should also be intelligent in order to help students by performing tasks like generating an appropriate study plan, giving recommendations of relevant material, doing intelligent searches, and offering personalized help.

The use of adaptive and intelligent user interfaces in a u-learning environment affects positively the students learning, by allowing them to study in a personalized environment at any hour, in any place, and from any computer. This way, students can center in the study of the subject, without having to lose time seeking for relevant material, or to depend on their computer skills in order to use the interface.

The aim of this paper is, first, to present our research on the main challenges of implementing an adaptive and intelligent user interface, and second, to report our first steps in implementing a u-learning environment for learning statistics. To this end, the rest of this work is organized as follows. Section 2 defines the basic concepts related to u-learning. Section 3 discusses the main problems that need to be solved to implement a platform of u-learning. Section 4 presents Statistics-to-Go, which is a project aimed to implement a m-learning platform for learning statistics. This is our first step toward the broader aim of implementing a u-learning platform. Finally, Section 5 presents our conclusions.

2 Basic Concepts of U-learning

This section defines the concepts of adaptive e-learning (Subsection 2.1), m-learning (Subsection 2.2), and u-learning (Subsection 2.3).

2.1 Adaptive E-learning

E-learning, also known as online learning, is defined as the instruction received by means of a CD-ROM, Internet, or intranet. E-learning can be synchronous or asynchronous [1, p. 10]. Synchronous e-learning is carried out in real time on a virtual classroom with an instructor. Thus, students and the instructor coincide in time though they are not necessarily at the same place. On the other hand, in asynchronous e-learning students and the instructor do not share the same time or place, and it is designed so students can learn at their own pace. Both formats allow the use of collaborative learning tools like wikis, discussion forums, and email.

Traditional e-learning offers the same service to all students without taking in account their previous knowledge, learning goals, communication skills, and preferences. Since the decade of 1970s, however, the impact of human differences on education were recognized [2]. That is, a "one-fits-all" education do not satisfy all students and then it is necessary to look for new teaching paradigms as, for example, adaptive e-learning.

Adaptive e-learning suggests a solution to this problem by adapting the instruction to the specific needs of each student. A more formal definition is that adaptive e-learning is an online teaching system that adapts selection and presentation of learning contents to the students in an individual way based on their previous knowledge, personal needs, learning style, and preferences [3, p. 24].

2.2 M-learning

M-learning is the acquisition of knowledge by means of some mobile device [4]. In this work, mobile device means cellular phones or personal digital assistants (PDAs).

The differences of m-learning with other types of learning, especially traditional e-learning, can be studied by considering both the technology involved and the educational experience. Regarding technology, m-learning differs by the use of portable equipment that allows students to access learning objects anytime and anywhere. Regarding the educational experience, Traxler [5] compares m-learning and e-learning using keywords. This way, m-learning is 'personal', 'spontaneous', 'opportunistic', 'informal', 'pervasive', 'private', 'context-aware', 'bite-sized', and 'portable', whereas e-learning is 'structured', 'media-rich', 'broadband', 'interactive', 'intelligent', and 'usable'. The same author notes that some of these distinctions can disappear as mobile technology advance, but properties as informality, mobility, and context will remain.

2.3 U-learning

U-learning is an adaptive e-learning system that allows students to learn anytime and anywhere in a personalized environment. In other words, u-learning is the combination of adaptive e-learning and m-learning. Following a similar formula found in [6–8] we define u-learning using the formula:

$$\text{u-learning} = \text{adaptive e-learning} + \text{m-learning} . \quad (1)$$

The benefits of u-learning are evident. It has the same ones that adaptive e-learning, that is, u-learning allows students to select their learning objectives and to apply their own learning style [3, 9, 10]. In addition, students can utilize any computer and study from any place in the world. The main disadvantage, for now, is the high cost of the mobile component and of the connection, particularly in cellular phones. Nevertheless, as technology advances it is predictable that prices will diminish.

3 Implementation of a U-learning Platform

This section describes the main problems that arise in implementing a u-learning platform (Subsection 3.1) and presents Bayesian networks, a mathematical tool, that can be used to solve these problems (Subsection 3.2).

3.1 Main Challenges

We have identified the main problems that need to be solved in order to implement a u-learning platform. Also, we have grouped these problems in two broad areas: infrastructure and software. In the infrastructure area it is necessary to

have a Web server and computers, both desktop and portable, with Internet access. In the software area, the design and implementation of an adaptative and intelligent user interface raises four principal problems: creation of the student model, generation of the interface content, adaptation of the interface content, and evaluation of the interface.

We think that infrastructure problems can be solved by means of financial resources. Here we notice some official initiatives which can make u-learning feasible. First, there are organisms, like the One-To-One Institute [11] and The Anytime Anywhere Learning Foundation [12], who facilitate computing equipment by promoting the "one to one" policy - one computer for each student. Second, there are governmental policies, like u-Japan [13] and u-Korea [14], aimed to facilitate Internet access to everyone. On the other hand, software problems need special attention since their solution involves aspects of computer science. For this motive, in this section we will detail the mentioned software problems.

The student model captures relevant information about the student, for example, previous knowledge level, learning goals, learning style, and available time for studying. The model must realize an initial evaluation of the student estimating his/her knowledge in the matter of study and his/her computer skills, and then evolve as the student's knowledge and skills change. Some problems that we identify in order the model can work in a suitable way are the following: 1) the model must learn rapidly to be useful from the first moment, 2) the model must work with incomplete, and possibly contradictory information, received from the student, and 3) the model must be able to evaluate the knowledge and skills of the student so it can evolve.

Content generation consists of producing a personalized user interface with the ability to realize intelligent tasks, like searches and recommendations, which redound in a more effective learning experience. The most important problems that it is necessary to solve are: 1) to establish a relevancy criterion to arrange the learning objects, 2) to generate a learning path depending on the student's learning goals and the available time for study, 3) to support a data base with the learning objects that the student has found relevant in order to recommend them to other students with similar goals, and 4) to discover and to recommend objects of learning that could be of interest for the student.

Content adaptation implies changing the user interface as the student's knowledge evolve and as the student move from one computer to another with, possibly, different capacities and screen size. To achieve these objectives it is necessary 1) to have a data base with the features, for example screen size, of every computational platform, and 2) to adapt the content to the capacities of the computer that is being used.

Finally, it is necessary to measure the performance of the user interface. This implies finding metrics to quantify the user satisfaction with the interface in order to compare among several design alternatives.

3.2 Bayesian Networks

One method of implementing an adaptive and intelligent user interface is utilizing *Bayesian networks*, which “are graphical structures for representing the probabilistic relationships among a large number of variables and doing probabilistic inference with those variables.” [15] This graphical structure is a directed acyclic graph, one can never return to the same node by following a sequence of directed edges. The graph has one node for each random variable in the associated probability distribution. Directed edges establish relationships among nodes. If there is an edge from node A to node B , then A is the “parent” of B and B is the “child” of A . This genealogical relation is often extended to identify the “ancestors” and “descendants” of a node.

Edges in a Bayesian network express the probabilistic dependencies between variables in a way consistent with the Markov condition: any node in a Bayesian network is conditionally independent of its nondescendants, given its parents [16, 17]. As a consequence, any Bayesian network specifies a canonical factorization of a full joint probability distribution into the product of local conditional distributions, one for each variable given its parents. That is, for a set of variables X_1, X_2, \dots, X_N , we can write:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N) = \prod_{i=1}^N P(x_i | Pa(X_i)) . \quad (2)$$

where $Pa(X_i)$ is the set of parents of X_i .

Bayesian networks can often be given a causal interpretation, an edge going from node A to node B indicates that A is a direct cause of B . This interpretation makes Bayesian networks particularly appealing for modeling high-level cognition. Once established the cause-effect relationships, there exist algorithms to make inferences.

Bayesian networks are a popular tool in applications that need to represent causal relationships and to reason with uncertainty [15, p. 619]. In e-learning, Bayesian networks can be used to determine the type of personalization with the intention of presenting a plan that optimizes the student’s learning process [3], to measure the knowledge of the student, as a part of the initial diagnostic [18], and to detect when the student has problems exploring the learning material in order to provide guidance to enhance this exploration [19].

4 Statistics-to-Go: Toward a U-learning Platform for Learning Statistics

As a first step toward implementing a u-learning platform, we decided to implement a m-learning platform. We chose statistics because that is the background of one of the authors. This section presents the *Statistics-to-Go* project, whose main objective is to devise tools for learning statistics in mobile devices. The section starts presenting some background information (Subsection 4.1) and

listing the main advantages of a m-learning platform (Subsection 4.2). Then, the project's objectives are presented (Subsection 4.3). Finally, the project's methodology is shown (Subsection 4.4).

4.1 Background

In the University of Sonora, students majoring in Mathematics, Physics, and Computer Science take a mandatory introductory course of statistics. The objective of the course is to teach students basic statistical tools and to familiarize them with statistical analysis using statistical software. These statistical tools run only in desktop or laptop computers, which implies students must be either in the computer laboratory or carrying their laptops.

On the other hand, taking in account that in 2009 there were almost 80 millions of cellular subscribers in Mexico [20] and that around the world 50% of cellular phones are Java-enabled [21], we want to know if cellular phones can be useful to solve statistics problems, helping students to learn statistics anytime and anywhere, and liberating them from carrying their laptops or of having to be in the computer laboratory.

This way, Statistics-to-Go project was proposed in February 2010 with the goal of producing tools for learning statistics in mobile devices using Java ME (micro edition). Java ME was selected because almost all mid- and high-range cellular phones are Java-enabled, and this means Java ME programs (called MIDlets) are portable across mobile operating systems.

4.2 Justification

The main advantages of a m-learning platform are:

- There are no time or place restrictions. Students can learn anytime and anywhere.
- It allows context-aware learning. Students can receive learning objects depending of their location.
- Students can communicate among them, or with the teacher, without physical contact.
- Teachers can design quizzes and give students instant feedback.
- Teachers can design learning objects with integrated video and audio.
- Video and pictures can be used as an alternative way of learning.

4.3 Objectives

General Objective

- To devise tools for learning statistics in mobile devices.

Specific Objectives

1. To research which kind of statistics tools are useful for students taking the introductory course of statistics at the University of Sonora.
2. To develop a prototype of these tools.
3. To deploy the tools to students taking the statistics course.
4. To evaluate the usability of the tools by the students.

4.4 Methodology

The Statistics-to-Go project is set to run from March 2010 to February 2011 and has the following activities:

1. To apply a survey to obtain the number of students at the University of Sonora who carry a Java-enabled cellular phone.
2. To apply a survey to former students and professors teaching the statistics course in order to know which statistics tools are convenient to develop in cellular phones.
3. To develop the tools.
4. To deploy the tools to students taking the statistics course.
5. To apply a survey to students taking the statistics course to know the usability of the tools and their impact in the students learning process.

5 Conclusions

U-learning benefits students because it allows them to study anytime and anywhere in a personalized way, independently of the computer used. In spite of these advantages, there are not many implementations of u-learning platforms for learning mathematics, much less for learning statistics. We also highlight the lack of suitable learning objects. We think this is due to the cost of the infrastructure needed. Nevertheless, it is predictable that the cost of technology gets affordable with time.

Our future work includes the implementation of a m-learning platform for learning statistics, the research of the application of Bayesian networks to the implementation of an adaptive and intelligent user interface in a u-learning platform, and the design of learning objects for statistics.

References

1. Clark, R., Mayer, R.: *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*. Pfeiffer, San Francisco, CA (2002)
2. Shute, V., Towle, B.: Adaptive E-Learning. *Educational Psychologist*, 38. 2. 105-114 (2003)

3. Tapia Moreno, F. J.: Modelo Bayesiano para la Optimización y Personalización del Proceso de Aprendizaje en Línea: Estudio Casuístico. PhD Thesis. Universidad de Las Palmas de Gran Canaria. Las Palmas de Gran Canaria, Spain (2007)
4. Traxler, J.: Defining mobile learning. In: International Conference Mobile Learning 2005. 261-266. Malta (2005)
5. Traxler J.: Defining, Discussing, and Evaluating Mobile Learning: The moving finger writes and having writ... . International Review of Research in Open and Distance Learning, 8, 2, 1-12 (2007)
6. Fraser, J.: u-Learning = e-Learning + m-Learning. <http://www.infotech.monash.edu.au/promotion/coolcampus/workshop/3rdworkshop/walkaboutlearning.pdf> (2005)
7. Ramón, O.: Del e-Learning al u-Learning: la liberación del aprendizaje. <http://sociedaddelainformacion.telefonica.es/jsp/articulos/detalle.jsp?elem=5162> (2007)
8. Wheeler, S.: U-Learning: Education for a Mobile Generation. <http://www2.plymouth.ac.uk/distancelearning/U-Learning.ppt> (2006)
9. Tapia, F., Galán, M., Ocón, A., Rubio, E.: Using Bayesian Networks in the Global Adaptive e-Learning Process. EUNIS 2005 (2005)
10. Tapia, F. J., López, C.A., Galán, M., Rubio, E.: Bayesian Model for Optimization Adaptive e-Learning Process. Journal of Emerging Technologies in Learning, 3, 2, 38-52 (2008)
11. One-To-One Institute. <http://one-to-oneinstitute.org/>
12. The Anytime Anywhere Learning Foundation, <http://www.aalf.org/>
13. u-Japan Policy, http://www.soumu.go.jp/menu_seisaku/ict/u-japan_en/index.html
14. u-Korea Master Plan. <http://www.unapcict.org/ecohub/resources/u-korea>
15. Neapolitan, R. E.: Learning Bayesian Networks. Prentice-Hall. Englewood Cliffs, NJ (2003)
16. Pearl, J.: Probabilistic reasoning in intelligent systems. Morgan Kaufmann. San Francisco. CA (1988)
17. Spirtes, P., Glymour, C., Schienens, R.: Causation prediction and search. Springer-Verlag. New York, NY (1993)
18. Conejo, R., Millán E., Pérez de la Cruz, J. L., Trella, M.: Modelado del alumno: un enfoque bayesiano. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 12, 50-58 (2001)
19. Bunt, A., Conati, C.: Probabilistic Student Modelling to Improve Exploratory Behaviour. Journal of User Modeling and User-Adapted Interaction, 13, 3, 269-309 (2002)
20. Mexico's mobile subscribers reach 79.4m, <http://www.totaltele.com/view.aspx?ID=445636>
21. Smartphone research, <http://uclue.com/?xq=1180>

Using the Software Process Improvement approach for defining a Methodology for Embedded Systems Development using the CMMI-DEV v1.2

García, I. and Herrera, A.

Postgraduate Department
Computer Science Faculty, Technological University of the Mixtec Region
Carretera a Acatlima km. 2.5. 69000 Oaxaca. Mexico.
{ ivan@mixteco.utm.mx; andrea@mixteco.utm.mx }

Abstract. Software process improvement holds a significant promise to reduce cycle times and provide greater value to all development activities involved in the software process development. While these methods appear to be well suited for embedded systems development, their use has not become an organized practice. In the same way as that of software development, the embedded systems development could be failing due a bad management in the development process. CMMI-DEV v1.2 is a process improvement maturity model that has been developed by the Software Engineering Institute at Carnegie Mellon. CMMI-DEV v1.2 defines "what" processes and activities need to be done and not "how" these processes and activities are done. In this paper we introduce the SPIES methodology that integrates the CMMI-DEV v1.2 Level 2 process areas to specify a process for developing embedded systems. This methodology incorporates the TSP principles to support the lack of management and improve the process specification. To illustrate this approach, we describe an experimental system in which it has been applied to develop and manage a traffic light system.

Keywords: Software process improvement, embedded systems, effective development process, improvement models, product focused improvement.

1. Introduction

Over the last 20 years, software's impact on embedded system functionality, as well as on the innovation and differentiation potential of new products, has grown rapidly. The consequence of this is an enormous increase in software complexity, shorter innovation cycle times, and an ever-growing demand for extrafunctional requirements (such as software safety, reliability, and timelines, for example) at affordable costs [15]. Moreover, these embedded systems' complexity is increasing, and the amount and variety of software in these products are growing. This creates a big challenge for embedded system development that was evident in the 2008 Embedded Market Survey, where respondents listed their three biggest concerns related to this task as

meeting schedules; the debugging process; and increased lines and complexity of the code. According to [19], of the same respondents a 59% also indicated that they are not using a formal development method or technique in their current embedded projects. As shown in Figure 1, meeting schedules is still the number one concern for developers. In fact, that concern actually increased by about 10% over last year. But, the most significant lesson is that meeting schedules is narrowly related to software development method, specifically to the planning process.

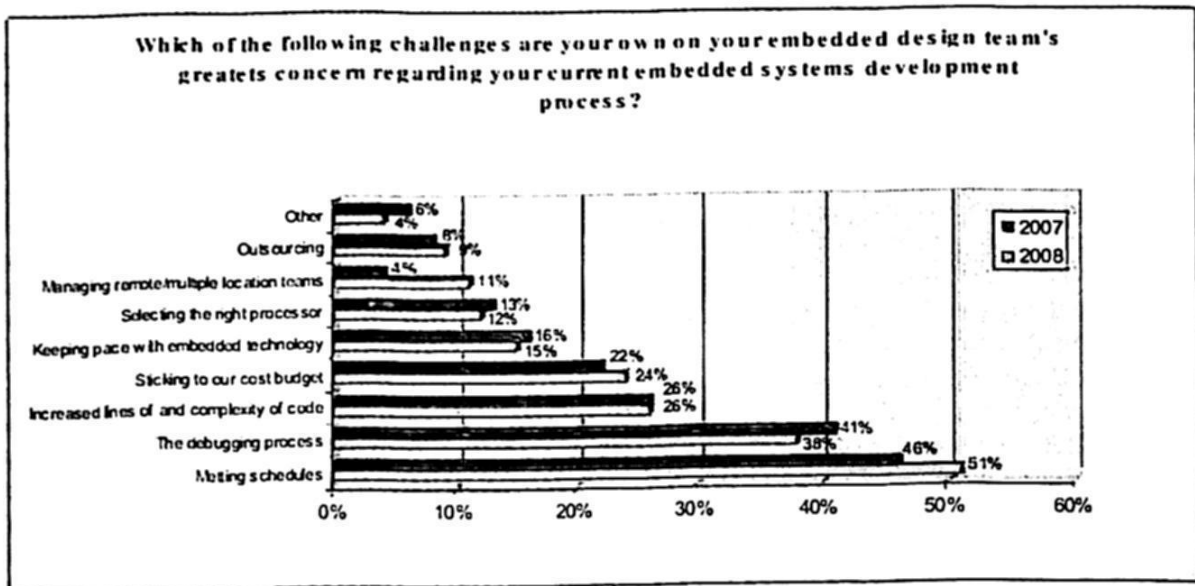


Fig. 1. Results of the 2008 embedded market study.

One potential reason for the poor adoption of development techniques is the gaps that exist when planning a detailed development lifecycle when the project begins. Traditionally, the gaps in the path of project planning and project monitoring and control to a solid product are overcome by experience and iteration. However for embedded systems, these gaps are more evident due to the need to obtain a realistic plan on affordable costs. This often results in embedded systems that exceed the established costs, overlap the schedule, and differ from the original functionality; necessitating extra time and resources to fulfill the initial requirements. Moreover, the limited availability of resources is preventing the introduction of new product features and applications, especially in areas where high-performance embedded systems are required [24]. Then, if there exists problems to develop conventional software products, are there problems related to develop the embedded software? In the same way, the commonly used software development process is a difficult task to perform when the developed software has to be run on an embedded system. Furthermore, due to cost, availability or developing reasons, the target system may not always be ready for the final phases (integration or testing time). These factors cause heavy bottlenecks when multiple developers work on the to-be-developed embedded system and need to concurrently access the development cycle for doing their tasks.

Just as we said, all these trends pose an urgent need for advanced embedded systems development techniques. However, and according to [14], the state of the art in embedded systems development is far behind other application areas. Graff et. al

studied seven European firms to determine the state of the practice in embedded software engineering [5]. One of the key findings in the study was that systems engineering decisions are largely being driven by hardware constraints, which then impact software efforts two stages later in the lifecycle when software requirements at the component level are developed. To optimize the timeliness, productivity, and quality of embedded systems development, companies must adapt software engineering technologies that are appropriate for specific situations. Unfortunately, the many available software development technologies do not take into account the specific needs of embedded systems development. In fact, developers do not tend to develop products with standardizations. For example, in Alcatel Shanghai Bell, the project scales have been different from several person/months to 150 persons/months. Project outlay is different from thousands USD to million USD, many of which were cancelled or failed during the development because of delay, over cost, out of control and customer unsatisfied. Problems being encountered in other enterprises can be found in this company too.

Our research aims to facilitate an alternative methodology for embedded systems development. We establish a formal development method oriented to embedded application fields which we called SPIES (Software Process Improvement for Embedded Systems). The SPIES methodology consists of three essential elements: activities, assets and tools. SPIES is supported by the CMMI-DEV v1.2 [27] to guide the whole project development.

The rest of this paper is organized as follows. Section 2 summarizes the related work of methodologies for developing embedded systems. Section 3 provides a brief description of CMMI-DEV v1.2. Section 4 formally describes the SPIES phases. A case study is presented in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

According to our literature review, much efforts have been paid to establish "what activities to implement" instead of "how to implement" these activities for embedded systems development. The current problem with embedded systems development is not a lack of standards or models, but rather a lack of effective strategy to successfully implement these standards or models. However, according to the MOOSE (Software Engineering Methodologies for Embedded Systems) project [18] we should consider other factors which could affect the project success:

- Coordinating all subprocesses (p.e. mechanical engineering, electrical engineering) to develop quality products is one of embedded system development's most challenging aspects.
- Systems engineering was mostly hardware driven – that is, from a mechanical or an electronic viewpoint. Consequently, software development started when hardware development was already at a stage where changes would be expensive.

Thus, since MOOSE results many approaches have been developed to incorporate formality to the embedded systems development process. In the following, we summarize the most significant work.

- The research by [29] was one of the first approaches that addressed conceptual models for relating together embedded systems' product and process characteristics. The purpose of the models was to describe in detail the characteristics of embedded systems for enhancing ISO 15504:1998 (SPICE) [10] conformant assessment methods to cover embedded systems' software process assessment. The analysis of embedded systems was based on the domain expertise of the PROFES (PROduct Focused improvement for Embedded Software processes) partners, a literature survey and assessments performed at several industrial sites.
- Years later, the enhancement of embedded products focused its efforts in the components reuse. The Koala component model [20], for example, was used for embedded software in consumer electronics devices, which allows late binding of reusable components with no additional overhead, but it does not take a development lifecycle into account and lacks a formal model.
- Later, PECOS [6] attempted to enable component-based technology for a certain class of embedded systems known as "field devices". The main features of this model are the data-flow-oriented programming style and the explicit incorporation of non-functional requirements. However, PECOS merely supports the non-functional properties of memory consumption and real-time, but does not provide any formal method supports, and only supports the data type interface.
- Since 2004, Miller and Smith have successfully used a prototype Test-Driven Development (TDD) [17] embedded system test framework, called Embedded-Unit, in their undergraduate classes and research work. In particular, the framework provides a stable, easy-to-learn environment through which students could solve problems generated at both the hardware and software levels. The TDD approach to support embedded system customer acceptance tests through Matlab-Fit and Embedded-FitNesse is best described as showing "potential".
- The Simplified Parallel Processes (SPP) [12] that provides a Software Process Improvement (SPI) solution referring CMMI [26] Level 2 and 3 for middle size enterprises. This research illustrates the development of a software processes management tool, called "Future", based on SPP to provide technical development specification for embedded systems. However, there doesn't exist a detailed explanation of SPP nor any real environmental validation or data of successful implementation.
- The REMES model for embedded systems by Secseleanu et. al [24] introduces a formal modeling and analysis of embedded resources such as storage, energy, communication, and computation. This model is a state-machine based behavioral language with support for hierarchical modeling, resource annotations, continuous time, and notions of explicit entry and exit points that make it suitable for component-based modeling of embedded systems. However, the analysis of REMES-based systems is only centered

- around a weighted sum in which the variables represent the amounts of consumed resources.
- The research by [14] presents a formal model for specification, verification, and composition of component-based embedded software which they called ESCM (Embedded Software Component Model). Li et. al describe how components are specified from the syntactical view, functional view, QoS view and synchronization view. The refinement rules for functionality, QoS, and synchronous behavior are defined for the verification purpose and a lightweight method is provided for the purpose of composition. This approach uses five contracts to give a formal specification of ESCM from the four separated levels view. ESCM's strict specification is required if one developer wants to safely reuse a component.
 - Sentilles et. al refine the component-based approaches through the development of Save-IDE [25]. Save-ID is an Integrated Development Environment for the development of component-based embedded systems that supports efficient development of dependable embedded systems using a dedicated component model, formal specification and analysis of component and system behaviors already in early development phases. In fact, the main contribution of Save-ID is related with its effort to establish a software development process, designated SaveCCT – SaveComp Component Technology, with three major phases: design, analysis and realization.

As Karsai et. al note in [13]: *"the development of software for embedded systems is difficult because these systems are part of a physical environment whose complex dynamics and timing requirements have to be satisfied"* Furthermore, the projects often lack an effective software development methodology [7]. Our contribution is more related with this last approach, defining and implementing a SPI effort to improve the development process establishing *effective practices* acquired from the good experiences of software commercial models, but covering the hardware elements implementation too.

3 The Capability Maturity Model Integration for Development

According to SEI, "CMMI is a process improvement maturity model for the development of products and services. It consists of best practices that address development and maintenance activities that cover the product lifecycle from conception through delivery and maintenance. This latest iteration of the model as represented herein, integrates bodies of knowledge that are essential for development and maintenance. These, however, have been addressed separately in the past, such as software engineering, systems engineering, hardware and design engineering, the engineering "-ilities," and acquisition" [26]. The prior designations of CMMI for systems engineering and software engineering (CMMI-SE/SW), are superseded by the title "CMMI for Development", to truly reflect the comprehensive integration of these bodies of knowledge and the application of the model within the organization.

CMMI-DEV v1.2 provides a comprehensive integrated solution for the development and maintenance activities applied to products and services.

The CMMI-DEV official report indicates that: "CMMI for Development v1.2 is a continuation and update of CMMI V1.1 and has been facilitated by the concept of CMMI "constellations" wherein a set of core components can be augmented by additional material to provide application-specific models with highly common content. CMMI-DEV is the first of such constellations and represents the development area of interest".

To improve software development practices, practitioners, projects, and organizations must move from ad hoc practices to explicit software development practices. Using CMMI-DEV v1.2 [27] and the IDEAL model [16] organizations can do just that. The IDEAL Model is used as the approach for this improvement process. The steps of the IDEAL Model are outlined below.

- *Initiating*: Laying the groundwork for a successful improvement effort.
- *Diagnosing*: Determining where you are relative to where you want to be.
- *Establishing*: Planning the specifics of how you will reach your destination.
- *Acting*: Doing the work according the plan.
- *Learning*: Learning from experience and improving your ability to adopt new technologies in the future.

Using the IDEAL model and CMMI-DEV v1.2, a process improvement team would improve its organization's software development practices. We believe that this approach can be applied with a high level of success to the embedded systems development. CMMI-DEV v1.2 is composed by 22 process areas which are divided in four categories and five maturity levels and/or capability levels to guide the process improvement. Table 1 shows that SPIES covers basically process areas of CMMI-DEV v1.2 Level 2, it means establishes a "managed process".

Table 1. CMMI-DEV v1.2 process areas for Level 2

Process area	Category	Maturity Level	Capability Level				
			1	2	3	4	5
Requirements Management	Engineering	2	Target Profile 2				
Project Planning	Project Management	2					
Project Monitoring and Control	Project Management	2					
Supplier Agreement Management	Project Management	2					
Measurement and Analysis	Support	2					

¹ The differences between v1.2 and v1.1 are explained in detail on the SEI Web site (www.sei.cmu.edu) in CMU/SEI 2006 TR-008.

Process and product quality assurance	Support	2			
Configuration Management	Support	2			

To achieve maturity level 2, all process areas assigned to maturity level 2 must achieve capability level 2 or higher. However, individual process areas can complete all practices and to achieve capability level 2. The relationship between process areas and SPIES' product life cycle are shown in the following section.

4 The SPI for Embedded Systems Methodology

According to [22], a methodology consists of a language to specify elements and relationships among system's components, and a whole process (activities, products, inputs, outputs, metrics, entry criteria, exit criteria, roles, and more), which indicates what parts of language to use, how to use them, and when to use them. SPIES specifies an integrated set of activities (adapted from CMMI-DEV v1.2) to guide developers during all development lifecycles to develop robust, capable and secure systems. Figure 2 shows the elements that interact to establish a SPI methodology for embedded systems development. The development process is designed as a top-down approach with an emphasis on continuous improvement as exposed in [28] [30] and [1].

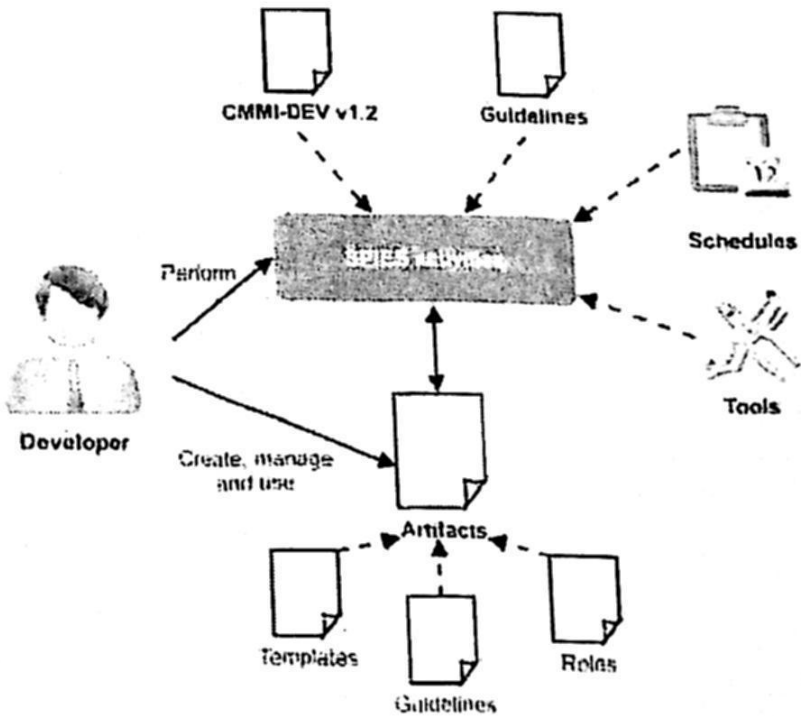


Fig. 2. The nature of SPIES methodology.

SPIES is organized by phases which are composed by more specific activities. The iterative approach of SPIES ensures that the development process be tested in each phase and not until the end of the project. As shown in Figure 2, our methodology uses an artifacts repository (or repository of assets) to introduce the process improvement in each phase. This repository contains templates for each activity (products), guidelines to perform activities, and role assignment. On the same way that CMMI-DEV v1.2, SPIES recommends for each activity (as possible) the use of an specific tool for project planning, estimation and effort process, requirements modeling, system validation, and more. For example, SPIES establish a set of activities to perform within the Requirements Specification phase using the Rhapsody Tool [23] from Telelogic (see Section 4 for a detailed explanation). SPIES use the basic idea of TSP [9] to establish a set of documents to guide developers in project management. The variation for incorporating this repository of knowledge to each phase ensures us high levels of success.

As we said, SPIES is extracted from process areas and specific practices in CMMI-DEV v1.2 level 2, which are simplified for embedded software demands. Figure 3 shows the relationship between CMMI-DEV v1.2 (at process areas level) and SPIES. Our methodology is composed by three layers and eight phases. Layers are dependent and related progresses. Every developer knows when to do an activity according the SPIES specification.

- The *management layer* provides the needed process to control the whole project development. This layer begins with the realization of a project plan and closes with the lessons learned in the final stage. These lessons are stored in the knowledge repository as needed effort, time, resources, people and more. This layer is composed by two process areas: Planning (PLA) and Product Continuous Improvement (PCI).
- The *engineering layer* provides the activities related to develop the complete system. The TSP artifacts are used in each SPIES phase (included in the Project Plan in previous layer) and determine when developers can begin the next phase through entry and exit criteria. The layer is composed by six process areas: Requirements Specification (RES), Product Design (PDS), Product Development (PDE), Product Integration (PIN), Product Validation (PVAL), and Product Delivery and Maintenance (PDM).
- The *support layer* provides help to achieve the expected quality through establishing activities for configuration and managing contracts and measuring them continuously. The layer is composed of three process areas: Configuration Management (CMA), Product and Process Quality Assurance (PPQ), Contract Management (CMG) and Measurement and Analysis (MAN).

Development process specification of Figure 3 belongs to CMMI-DEV v1.2' engineering processes category (Requirements Specification, Product Design, Product Integration, and Product Validation). Therefore, concrete technical development specification (Product Development) is necessary for applications. Specification defines phases of processes, activities and subactivities, entry and exit criteria, measures, etc; templates that illustrate formats for different documents, and how to

fulfill them; guidelines that specify how to realize and tailor specification and phases. Thus, the development process for embedded systems includes technical development specifications and their templates, guidelines, and more assets. In addition, hardware related processes are emphasized here.

The addition of the PLA process for embedded systems improve the conventional methods commonly used for embedded systems development in industrial environments; (for example the Model-Driven Architecture [21]; standard IEC 61508 [11]; the UML-based Rapid Object-Oriented Process for Embedded Systems, which is based on the spiral process model [2]; and different approaches based on a V-Model [3]) to establish realistic plans.

Across the eight phases of SPIES the information flows in the form of processes assets. The knowledge repository manages all information about the project (from project plans and contracts to improvement information – effort, time, and more) to continuously improve the embedded system development process.

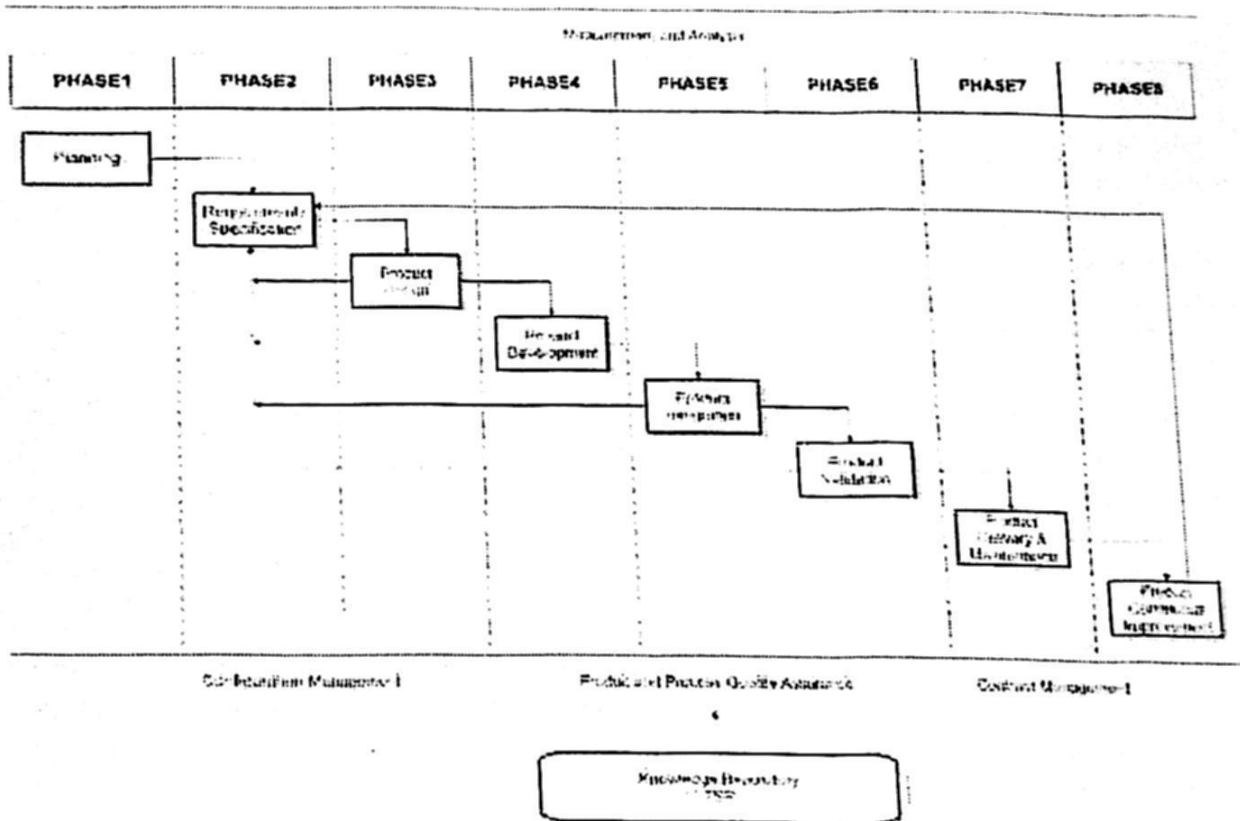


Fig. 3. SPIES phases and process.

According to Hansen et. al [8] “*model can be found to solve problems but there is a gap between theory and practice*”. Thus, we think that a suite of general SPI solutions can be outlined for different enterprises which look for establishing a “quality development process”. We try to unify “theory” and “practice” through the use of a knowledge repository using the TSP practices and templates as support and guiding developers through the CMMI-DEV v1.2 effective practices.

4.1 The SPIES structure

The SPIES structure enables developers to easily tailor specifications for their own necessities. SPIES has, from PHASE1 to PHASE8, 16 process areas and about 25 templates. When applying SPIES it is suggested modifying properly according to concrete situations (such as measures, capability levels, entry and exit criteria, etc.). In SPIES step by step is very important through an iterative approach. Figure 4 provides a detailed explanation of SPIES processes (CMA, PPQ, CMG, PPQ, and MAN are not shown because their activities are implicit in the other 11 processes).

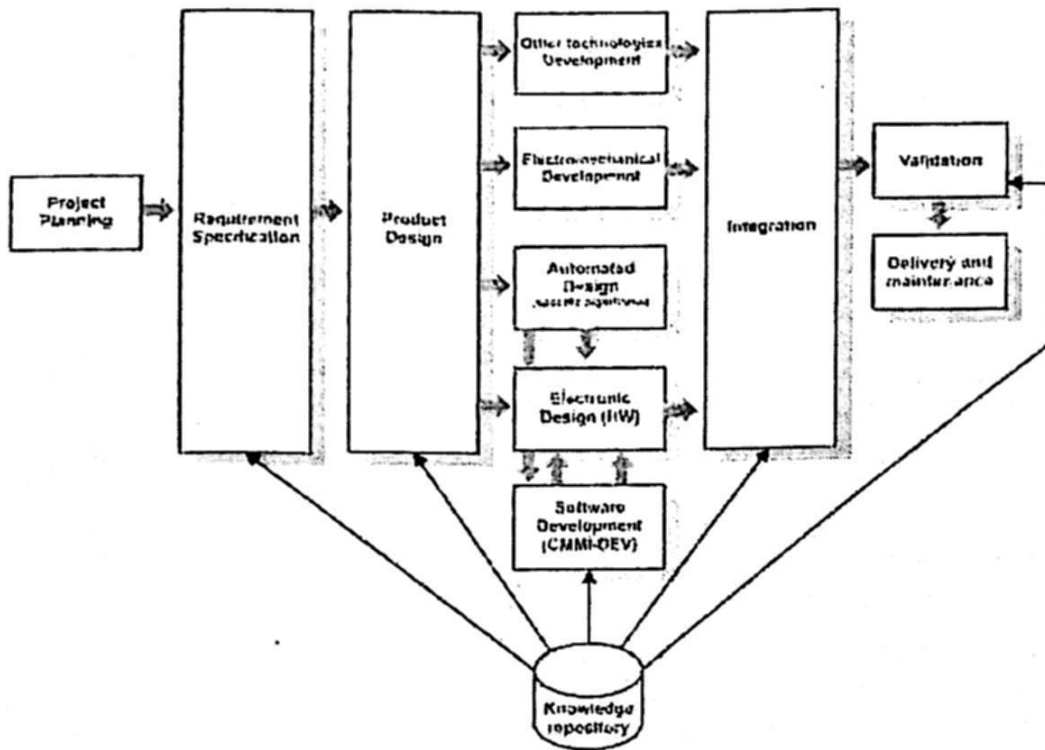


Fig. 4. SPIES process areas.

The contents of SPIES process can not be modified, while the modification of the previous projects assets (as documents, templates, guidelines, estimations, etc.) is enabled. The TSP model provides all templates from project initialization (planning) to project closing (delivery and maintenance). We provide a brief description of SPIES processes:

- **Project planning.** This process provides guide to establish and maintain plans that define project activities. Planning includes estimating the attributes of products and tasks, determining the resources needed, producing a schedule, identifying and analyzing project risks, and considerations for choosing the right microprocessor for an embedded application. Iterating through these SPIES' activities may be necessary to establish the project plan.
- **Requirement specification.** This process provides activities needed for requirements development and requirements management. Depending on the modeling languages used to describe requirements, developers can use

graphical languages such as UML or the Systems Modeling Language (SysML) to model requirements. SPIES recommends activities to use Rhapsody Tool from Telelogic.

- **Product design.** This process provides activities for creating a system's effective design based on the previous requirements and focused on two issues: functional design and architecture design. Functional designs cover an embedded system's functionalities without considering any technical implementation details. SPIES indicates to developers how to define the system's architecture on the basis on the requirements and functional design. The obtained architecture definition consists of various views covering the architecture's different aspects.
- **Other technologies development.** The quality enables performing the objective or subjective evaluation of the performed functionality. It is possible that the embedded system should use other technologies which affect the planned functionality. Thus, this process provides practices to develop embedded systems without technological dependencies.
- **Electromechanical development.** This process provides activities to enable developers to manage and synchronize information in elements of electrical, mechanical and software designs. SPIES performs automatic controlled reviews to identify potential problems among disciplines and use collaboration tools to quickly solve incidences in the process.
- **Automated design.** This process provides practices to integrate all system's components into an unified model. The process considers third-party requirements during the design of an embedded system and its software. A whole simulation is carried out.
- **Electronic design.** This process contains activities to use tools for designing and producing electronic systems ranging from printed circuit boards (PCBs) to integrated circuits. This is sometimes referred to as ECAD (electronic computer-aided design) or just CAD.
- **Software development.** This process contains activities to refine or extend platform-independent models in the platform-specific design, to support efficient code generation for the target execution platform. SPIES support the concurrent development with the hardware and other components of the embedded system through design and development phases.
- **Integration.** This process provides activities to integrate the components in several steps before the embedded system is tested. The integration phase of the development cycle must have special tools and methods to manage the complexity. The process of integrating embedded software and hardware is an exercise in debugging and discovery.
- **Validation.** The last validation phase may include extensive field testing and validation, before the embedded system is ready for delivery and maintenance.
- **Delivery and maintenance.** The majority of embedded system designers (around 60 percent) maintain and upgrades existing products, rather than design new products. Developers should use activities to use the existing documentation and the embedded system to understand the original design well enough to maintain and improve it. This process requires tools that are

especially tailored to reverse engineering and rapidly facilitating “what if” scenarios.

- **Measurement and analysis.** The measurement process is a prerequisite for all previous processes and for the successful process improvement. In this context, the SPIES measurement should be used for two purposes: evaluating conformance of an embedded system to the quality specification, and evaluating process-system relationships. This conceptual methodology includes measurement explicitly and applies it for these two purposes. SPIES defines criteria to which an embedded system development focused on SPI needs to comply. These criteria are depicted in Figure 5, together with an overview of the relationships between the criteria and the proposed layers.

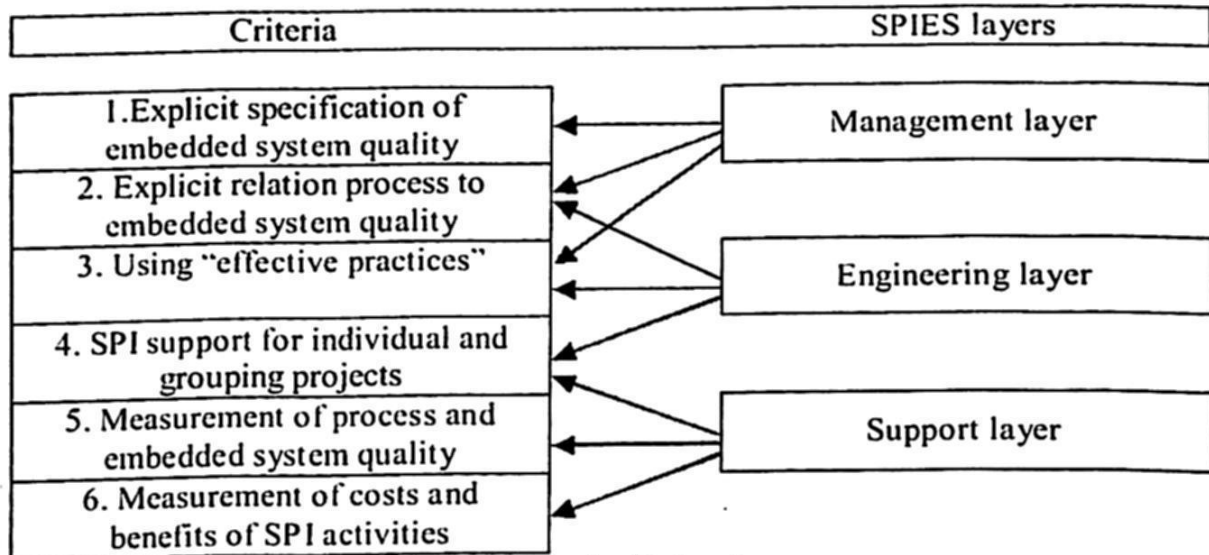


Fig. 5. Relation between SPIES criteria for embedded systems.

5 Testing the SPIES methodology

To evaluate the applicability of our methodology, we have successfully implemented a prototype with SPIES development, called the traffic light system, in our undergraduate classes and research work. During the original course, we gave students a complete course of embedded systems development. The traffic light system was designed to be used inside the faculty campus and includes three components: the sensor component for detecting vehicles and pedestrians, a Front Panel Display (FPD) to automatically configure the system, and the internal component to manage the traffic. It is important to say that as we are talking about a University environment, there is no such traffic as a common street. As illustrated in Figure 6, the traffic management component includes five modes: safe intersection mode, evening low volume mode, responsive cycle mode, fixed cycle time mode, and adaptive mode. The sensor can detect vehicles (priority vehicle and emergency vehicle) and pedestrians. This figure was obtained by RES activities related to Rhapsody tool.

We will briefly explain how the SPIES methodology is used. Our methodology begins with the planning process. Our students fulfilled the "Project plan template" to start the embedded system development. Managing embedded projects can be a conflict in disciplines, with the need to foster engineering creativity while wrapping the students in enough process to keep them on track. SPIES grouping contains effective practices for creating a project plan, tracking the plan against reality, managing contracts and delivering a project on time and budget. Figure 7 shows an example of two SPIES templates to generate a project plan.

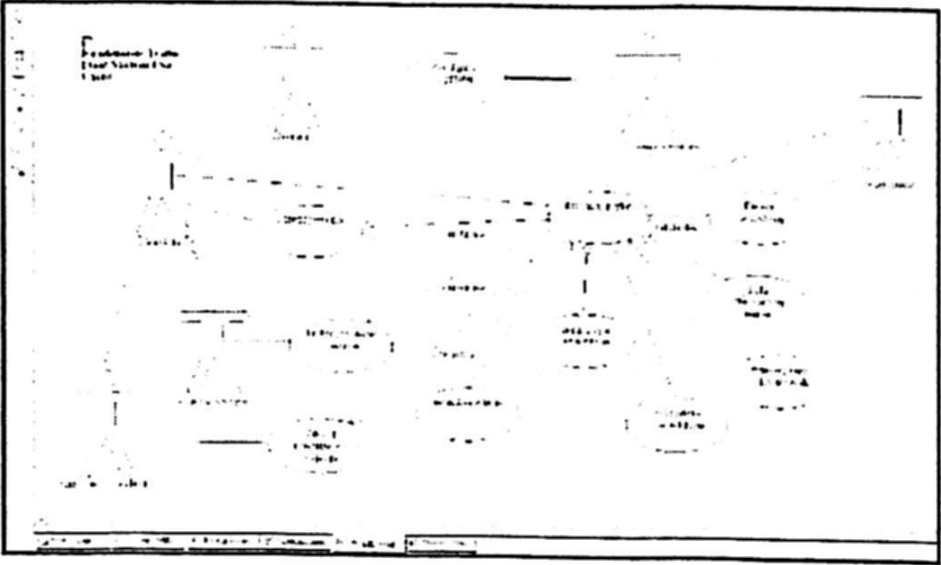


Fig. 6. The traffic light system.

SPIES Template for Project Planning (PROJ)

Name _____ Cycle _____ Date _____
Team _____ Week _____

SPIES for Summarized Plan: SUMP

Name _____ Date _____
Project Name _____ Cycle _____
Product _____

Product Size	Planned	Actual
High level design pages (UML)		
Low level design pages (C/C++)		
Requirements pages (REQ)		
Test cases (TCT)		
LOC: Base (module)		
LOC: Header		
LOC: Modified		
LOC: Added		
LOC: Revised		
LOC: Total / new changes		
LOC: Total		
LOC: Total / new revised		

Time per Phase	Planned	Actual	% Actual
Management and W/requirements			
Analysis and strategy			
Planning			
Requirements			

Fig. 7. An example of project plan generated by SPIES.

Table 2. An example of SPIES activities for RES process

RES 1.1	Identify and describe ideas for new system
RES 1.2	Prepare preliminary description of the new system
RES 1.3	Establish system requirements
RES 1.4	Identify interface requirements
RES 1.5	Establish operational and functional scenarios
RES 1.6	Establish a definition of required functionality
RES 1.7	Validate requirements
RES 1.8	Manage requirements changes
RES 1.9	Maintain bidirectional traceability of requirements

Student responsibility for collecting information and feedback from the experience plays an important role in this phase. All the collected information is updated in the knowledge repository to increase the set of effective practices and relate it to future projects. The role of RES is emphasized in the Product Design phase, where the system level architecture of the system implementation plan is refined with specifications for the software, electronic hardware, mechanical hardware, and more. RES 1.2, 1.3, 1.4 and 1.5 activities are supported by additional activities related to Rhapsody Tool. SPIES includes these activities to establish a formal process for obtaining and understanding the embedded system requirements. Figure 8 illustrates a set of general requirements, based on SPIES activities for modeling on Rhapsody, for our experimental project.

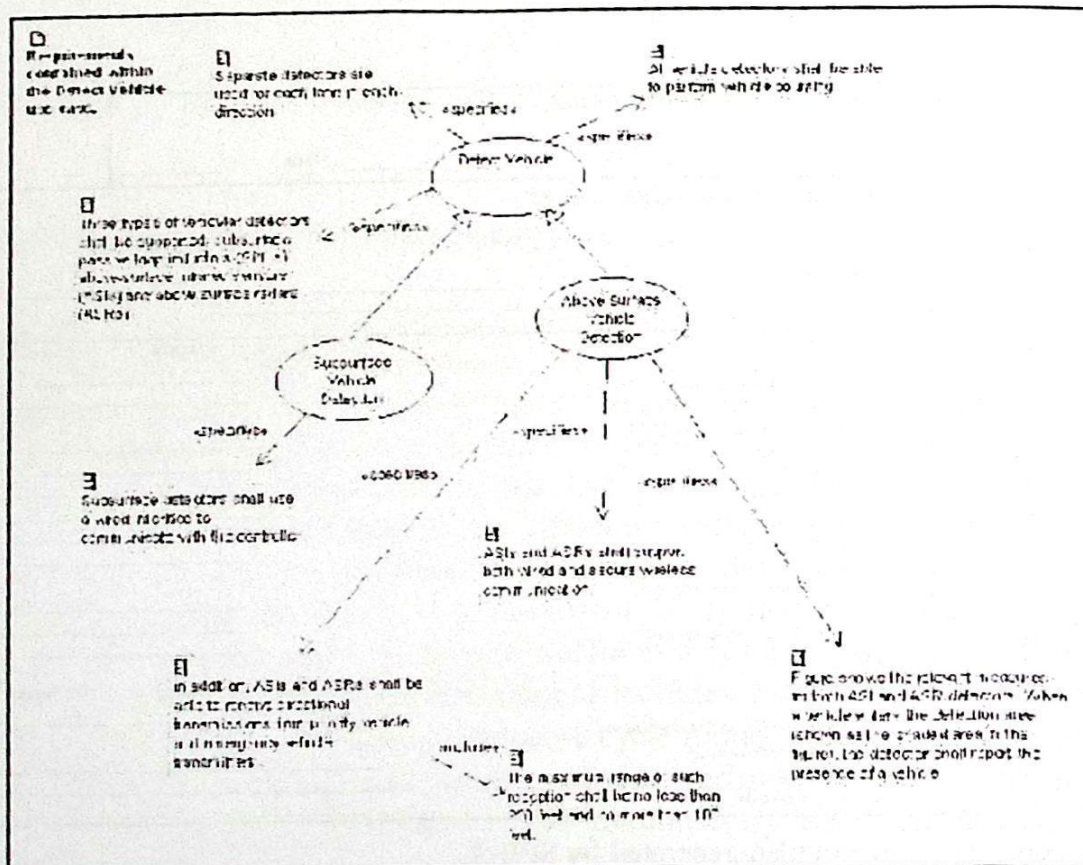


Fig. 8. SPIES in requirements modeled with Rhapsody.

In the traffic light system described above, we only give a general explanation of how SPIES works in specific process areas (specifically RES) and omit other processes specification which are given in Section 4.1 because of the length. A photograph of the traffic light system can be seen in Figure 9. It is developed with our SPIES methodology and software architecture and was developed in our group in collaboration with postgraduate students and research partners.

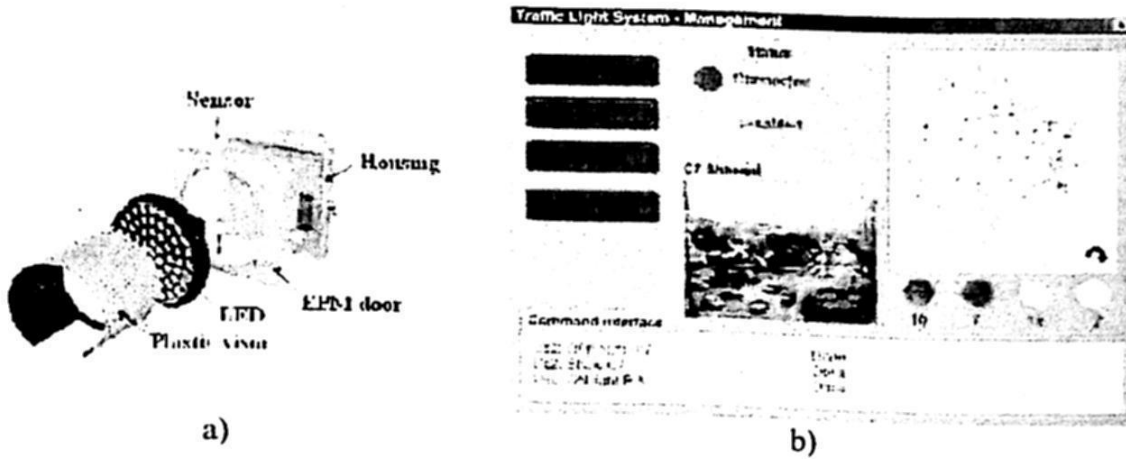


Fig. 9. a) The traffic light component. b) The management traffic light system.

Fortunately, using the engineering layer (discussed earlier in Section 4), we can shield the embedded software from hardware dependencies. Hence, the embedded software could be designed by following the SPIES activities and by modeling and simulation in Matlab/Simulink and Rhapsody tool, while abstracting from hardware details. Finally, the code is generated in C++ and deployed.

5 Conclusions and Future Research

The methods used, tools and techniques for embedded systems development do not only vary between companies and research institutes, but also within companies and universities themselves. Frequently, there is a general high-level approach present, but not much is standardized on a more detailed level. Thus different projects often use different tools and notations. Besides, a relevant factor mentioned in [4] when applying SPI approach to embedded systems development is the limited management support that disenables the successful factor. We think that the combination of two quality models (CMMI-DEV v1.2 and TSP) to improve the development cycle can support a formal process to avoid this lack of management. It is true that SPI depends on a company's capabilities (human, infrastructure and resources) and business strategies. But, the process development relies on standardization management too. CMMI-DEV v1.2 concentrates on project management and pay attention to technical course. We think that the standardization on technical processes for embedded systems is a crucial element to improve the quality of the final product.

In this paper, we presented a methodology for developing embedded systems using the SPI approach. We define the SPIES methodology and phases for the development process specification for embedded systems supported by TSP model. The SPIES methodology is based on effective practices and cheap compared with buying tools separately (in fact our methodology is Open Source). We briefly show experience about the implementation of a traffic light system using SPIES methodology. One of the observed disadvantages of SPIES is related to the previous knowledge on quality models that could be difficult to use. The repository knowledge can solve this problem through feedback from users.

As future work, we will refine effective practices through experimentation giving a formal architecture style for the construction of an automated software that implements the practices reflected in SPIES.

We will implement our methodology in enterprises, and modify it according to users' lessons learned. Finally, more experiments in academia will be defined and carried out at one or more of the industrial partners to test SPIES in real-world environments.

References

1. Basili, V., McGarry, F., Pajerski, R. & Zelkowitz, M. "Lessons learned from 25 years of process improvement: The Rise and Fall of the NASA Software Engineering Laboratory" *Proc. of the 24th International Conference on Software Engineering* (ICSE 2002), pp. 69-79, 2002.
2. Boehm, B. "Guidelines for Verifying and Validating Software Requirements and Design Specification" *Proc. of the European Conference of Applied Information Technology* (Euro IFIP), North-Holland, pp. 711-719, 1979.
3. Boehm, B. "A Spiral Model of Software Development and Enhancement" *Computer*, 21(5):61-72, 1988.
4. Graaf, B., Lormans, M. & Toetenel, H. "Software Technologies for Embedded Systems: An Industry Inventory" *Proc. of the 4th International Conference on Product Focused Software Process Improvement* (PROFES 2002), LNCS 2559, pp. 453-465, 2002.
5. Graaf, B., Lormans, M. & Toetenel, H. "Embedded Software Engineering: The State of the Practice" *IEEE Software*, 20(6): 61-69, 2003.
6. Genßler, T., Christoph, A., Winter, M., Nierstrasz, O., Ducasse, S., Wuyts, R., Arévalo, G., Schönhage, B., Müller, P. & Stich, C. "Components for embedded software: the PECOS approach" *Proc. of the International Conference on Compilers, Architecture and Synthesis for Embedded Systems* (CASES 2002), ACM press, pp. 19-26, 2002.
7. Greene, B. "Agile methods applied to embedded firmware development" *Proc. of the Agile Development Conference* (ADC 2004), pp. 71-77, 2004.
8. Hansen, B., Rose, J. & Tjørnehøj, G. "Prescription, description, reflection: the Shape of the Software Process Improvement Field" *International Journal of Information Management*, 24(6): 457-472, December 2004.

9. Humphrey, W. "Introduction to Team Software Process", Addison-Wesley, Reading, MA, 2000.
10. ISO/IEC TR 15504:1998(E). *Information Technology – Software Process Assessments. Parts 1-9*. International Organization for Standardization: Geneva, 1998.
11. International Electrotechnical Commission. IEC 61508, Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems, 1998.
12. Jun, D., Rui, L. & Yi-min, H. "Software Processes Improvement and Specifications for Embedded Systems" *Proc. of the 5th ACIS International Conference on Software Engineering Research, Management & Applications (SERA 2007)*, IEEE Computer Society, pp.13-18, 2007.
13. Karsai, G., Sztipanovits, J., Ledeczi, A., "Model-integrated development of embedded software" *Proceedings of the IEEE*, 91(1): 145-164, 2003.
14. Li, C., Zhou, X., Dong, Y. & Yu, Z. "A Formal Model for Component-Bases Embedded Software Development" *Proc. of the International Conference on Embedded Software and Systems (ICSS 2009)*, IEEE Computer Society, pp. 19-23, 2009.
15. Liggesmeyer, P. & Trapp, M. "Trends in Embedded Software Engineering" *IEEE Software*, 26(3): 19-25, 2009.
16. McFeeley, B. "IDEAL: A User's Guide for Software Process Improvement" *CMU/SEI-96-HB-001*, Software Engineering Institute, Carnegie Mellon University, 1996.
17. Miller, J. & Smith, M. R. "A TDD Approach to Introducing Students to Embedded Programming" *Proc. of the 12th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2007)*, ACM Press, pp. 33-37, 2007.
18. MOOSE project. Available at: <http://www.mooseproject.org/> [On line]. 2009.
19. Nass, R. "An insider's view of the 2008 Embedded Market Study" Available at: <http://www.embedded.com/design/210200580>. January, 2008.
20. Ommering, R. V., Linden, F. V. D., Kramer, J. & Magee, J. "The Koala Component Model for Consumer Electronics Software" *IEEE Computer*, 33(3): 78-85, 2000.
21. Petrasch, R. & Meimberg, O. *Model Driven Architecture – Eine praxisorientierte Einführung in die MDA*, Heidelberg, Germany: dpunkt, 2006.
22. Powel, B. "Real-time UML Workshop for Embedded Systems" Elsevier, Boston USA. 2007.
23. Powel, B. "The Telelogic Harmony/ESW Process for Real-Time and Embedded Development" Telelogic White Paper. Telelogic, 2008.
24. Secelcanu, C., Vulgarakis, A. & Petterson, P. "REMES: A Resource Model for Embedded Systems" *Proc. of the 14th IEEE International Conference on Engineering of Complex Computer Systems (ICECCS 2009)*, IEEE Computer Society, pp. 84-94, 2009.
25. Sentilles, S., Petterson, A., Nyström, D., Nolte, T., Petterson, P. & Crnkovic, I. "Save-IDE – A Tool for Design, Analysis and Implementation of Component-Based Embedded Systems" *Proc. of the 31st International Conference on Software Engineering (ICSE 2009)*, IEEE Computer Society, pp. 607-610, 2009.

26. Software Engineering Institute. *CMMI for Systems Engineering, Software Engineering, Integrated Product and Process Development, and Supplier Sourcing (CMMI-SE/SW/PPD/SS, V1.1)*. Continuous Representation. CMU/SEI-2002-TR-011, Software Engineering Institute, Carnegie Mellon University. 2002.
27. Software Engineering Institute. *CMMI for Development (CMMI-DEV V1.2)*. CMU/SEI-2006 TR-008, Software Engineering Institute, Carnegie Mellon University. 2006.
28. Solingen, R. V. "Product Focused Software Process Improvement – SPI in the embedded software domain" Eindhoven University of Technology, The Netherlands. 2000.
29. Taramaa, J., Khurana, M., Kuvaja, P., Lehtonen, J., Oivo, M., & Seppänen, V. "Product-Based Software Process Improvement for Embedded Systems" *Proc. of the 24th Euromicro Conference*, IEEE Computer Society, pp. 905-912, 1998.
30. Trienekens, J., Kusters, R. & Solingen, R. V. "Product Focused Software Process Improvement: Concepts and Experiences from Industry" *Software Quality Journal*, 9(4): 269-281, 2001.

Population Coding and SpikeProp Hardware Accelerator for Spiking Neural Networks

Marco Aurelio Nuño-Maganda¹, Cesar Torres-Huitzil², and Miguel Arias-Estrada³

¹ Universidad Politécnica de Victoria (UPV), Av. Nuevas Tecnologías S/N, Parque Científico y Tecnológico de Tamaulipas, C.P. 87137, Ciudad Victoria, Tamaulipas, México

² CINVESTAV-Tamaulipas, Information Technology Laboratory, Parque Científico y Tecnológico de Tamaulipas, C.P. 87137, Ciudad Victoria, Tamaulipas, México

³ National Institute for Astrophysics, Optics and Electronics (INAOE), Luis Enrique Erro #1, San Andrés Cholula, C.P. 72840, Puebla, México

Abstract. Spiking Neural Networks (SNNs) have become an important research topic due to new discoveries and advances in neurophysiology, which suggest that information among neurons is coded, interchanged and processed via pulses or spikes. Two important aspects of SNNs are coding and learning as means to represent information and acquire knowledge by experience, respectively. The SpikeProp algorithm has been proposed as a learning algorithm for SNNs with good results in classification and pattern discrimination. SpikeProp algorithm requires a coding scheme called Gaussian Receptive Fields (GRFs), which converts real data to firing times. In this paper we explore the feasibility of using FPGAs for developing specialized hardware modules for GRF coding and for large scale simulations of SNNs. Modularized processing elements were designed to evaluate different implementation tradeoffs and to promote scalability of the model to larger FPGAs. Results and performance statistics are presented, as well as a discussion of implementation trade-offs using pattern classification problems as a case study.

1 Introduction

Spiking Neural Networks (SNNs), classified as neural models of third generation [1], have become an important research area due to its biological plausibility. Roughly speaking, the behavior of ideal spiking neurons is described as follows: A typical neuron can be divided into three functionally distinct parts, called dendrites, soma and axon. The dendrites play the role of the “input device” that collects signals from other neurons and transmit them to the soma. The soma is the “central processing unit” that performs a nonlinear processing step. If the total input exceeds a certain threshold, then an output signal is generated. Then, the output signal is converted by the “output device”, the axon, which delivers the signal to other neurons. The biological neuronal signals consist of short electrical pulses. The pulses, the so-called action potentials or spikes, have an amplitude of about 100 mV and typically a duration of 1-2 ms [2]. It is recognized

that SNNs are capable of exploiting time in a sophisticated manners a resource for information coding and computation. Learning in traditional, artificial neural networks is usually performed by gradient ascendant/descendant techniques to find the network parameters to perform a given task. For SNNs these techniques are extrapolated and the SpikeProp algorithm, has been proposed for learning [3] [4] [5] [6] [7] for neurons modeled by the Spike Response Model (SRM) [8].

The motivation of digital implementations of SNNs is as diverse as the background of the researchers, but most implementation look for performance speedup and/or a large scale simulations. Spiking models are less hardware-greedy than classical models, and accuracy results compared to classical neural networks are similar [9]. Due to programmability, digital hardware offers a high degree of flexibility and provides a platform for simulations on neuron level as well as on network level [10]. In [11], several hardware platforms used for simulation of SNNs are reported, where the network sizes range from 8K to 512K neurons. The presented results concluded that only supercomputers can cope with the computer processing demands of these type of networks, however other alternative exist, such as FPGA-based computing, for such kind of processing.

This paper is divided in 5 parts. Section 2 presents an overview of coding schemes and recall phase of feedforward SNNs. Section 3 describes the proposed architecture for each stage of the SNN processing. In Section 4, experimental results about performance of hardware blocks are presented, with a tradeoff discussion. Finally, in section 5, conclusion and future work are presented.

2 Background

2.1 Population Coding using Gaussian Receptive Fields (GRFs)

The application of GRFs for supervised learning was introduced in [3], where a set of analog variables is fed into the GRFs to convert them in pulse streaming suited for the SNNs processing and learning. A Gaussian function is defined by:

$$f(x) = ae^{\frac{(x-b)^2}{2\sigma^2}} \quad (1)$$

for real constants $a > 0$, the height of the Gaussian peak, $b > 0$, the position of the center of the peak and $\sigma > 0$ controls the width.

A real input value is encoded by an array of receptive fields. The range of the data is first calculated, and then, each input feature is encoded with a population of neurons that cover the whole data range. For a variable with a range $[I_{min}^n, \dots, I_{max}^n]$, a set of m GRF neurons are used. The center b_i of each GRF neuron is determined by:

$$b_i = I_{min} + \frac{1}{m-2} \frac{2i-3}{(I_{max} - I_{min})}, \quad (2)$$

And the width σ of each GRF neuron i is determined by:

$$\sigma = \frac{1}{(m-2)} \frac{1}{\beta(I_{max} - I_{min})}, \quad (3)$$

where the proposed value for β belongs to the range $[1, 2]$.

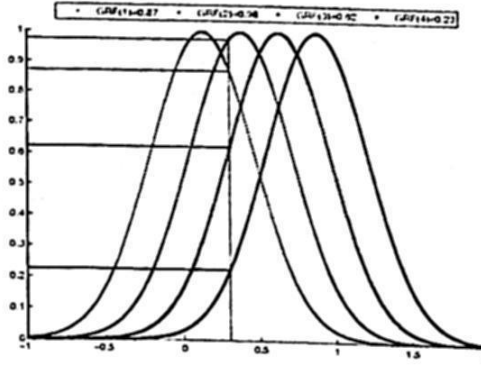


Fig. 1. Example of GRFs coding

The process of coding an analog variable is shown in figure 1, where the total number of GRFs is four, and the overlapped GRFs are plotted together. The value to be coded is 0.3 (shown as a vertical line), and the evaluation of that value in each GRF is shown by a horizontal line.

2.2 Spiking processing in recall phase

A neuron j , having a set Γ_j of immediate predecessors (“pre-synaptic neurons”), receives a set of spikes with firing times $t_i, i \in \Gamma_j$. Any neuron generates at most one spike during the simulation interval, and fires when the internal state variable reaches a threshold ϑ . The dynamics of the state variable $x(t)$ are determined by the input spikes, whose impact is described by the spike-response function $\Delta(t)$ weighted by the synaptic efficacy (“weight”) w_{ij} :

$$x_j(t) = \sum_{i \in \Gamma_j} w_{ij} \varepsilon(t - t_i) \quad (4)$$

A network consists of a fixed number of m synaptic terminals (or connections), where each terminal serves as a sub-connection that is associated with a different delay d and weight w . The unweighted contribution of a single synaptic terminal k to the state variable y is given by:

$$y_i^k = \varepsilon(t - t_i - d^k) \quad (5)$$

where $\varepsilon(t)$ is the spike response function, with $\varepsilon(t) = 0$ for $t < 0$. The time t_i is the firing time of a previous neuron i , and d^k is the delay associated with the synaptic terminal k . The spike response function is given by:

$$\varepsilon(t) = \frac{t}{\tau} e^{(1-\frac{t}{\tau})} \quad (6)$$

Extending (4) to include multiple synapses per connection and, substituting in (5) the state variable x_j of neuron j receiving input from all neurons i , the network can be described as the weighted sum of the pre-synaptic contributions:

$$x_j(t) = \sum_{i \in I_j} \sum_{k=1}^m (w_{ij}^k y_i^k(t)) \quad (7)$$

The target of learning through the SpikeProp algorithm is to get a set of firing times t_j^k , at the output neurons $j \in J$ for a given set of input patterns $P[t_1...t_h]$, where $P[t_1...t_h]$ denotes a single input pattern described by single firing times for each neuron $h \in H$. The error-function is defined by:

$$E = \frac{1}{2} \sum_{j \in J} (t_j^a - t_j^d)^2 \quad (8)$$

where t_j^d are the desired firing times and t_j^a are the actual firing times. The details of the SpikeProp algorithm are out of the scope of this paper, but it is being considered for further on-chip implementation. See [3] for further details on SpikeProp.

3 Proposed Architecture

The dataflow for the architecture is shown in figure 2. The input data set is organized in rows and columns. The columns are the dataset attributes, while the rows are the dataset samples. The class column is stored in a separated memory region. The input dataset is passed through the GRFs, which main function consists of codifying the input data into firing times. Once this transformation has been carried out, the firing times are passed to the SNN module, which computes the output as firing times. Then, the output firing times are passed to a class decoder, which obtains the class assigned by the network to the input pattern. Finally, the network performance can be evaluated, comparing the class obtained by the network with the original class assigned to that pattern. Additionally, a hardware-based learning module can be implemented for adjusting weights and delays for learning from a dataset for a specific application.

3.1 Architectural Overview

The proposed hardware for carrying out the dataflow previously explained is shown in figure ?? . A system with 2 data buses is proposed. The Global Data Bus connects external memory elements with internal routers. The internal routers send data to each one of the processing elements. The processing elements read from the bus their input data, process them and output the results on other data bus, which feed an output router. The output router can send data to memory, or to other router (the input router of other process).

The control bus contains all the control signals generated by the global control unit. The control unit generates the synchronization signals required for each

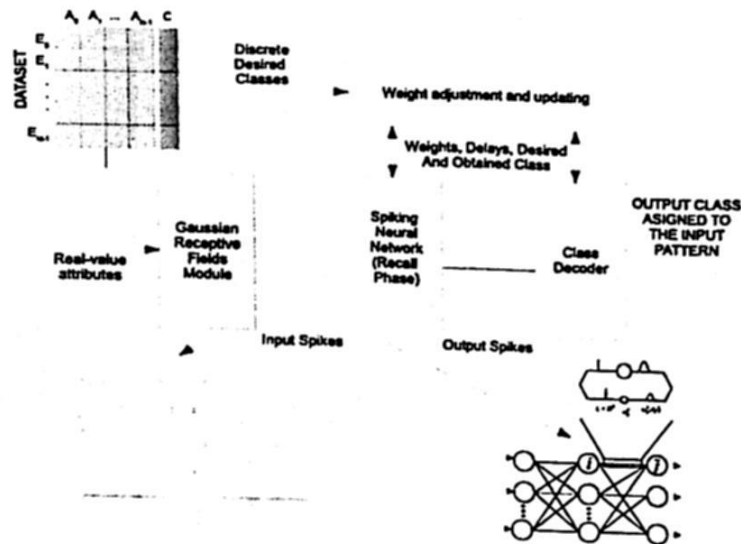


Fig. 2. Architectural dataflow

module in the proposed architecture. There are two type of processors. The first one is the Gaussian Receptive Fields Processor (GRPF), which transform the input data into firing times. The second one is the Spiking Neural Layer Processor (SNLP), which obtains the network output for a given input firing times pattern. These processors are explained more in detail in this section.

The proposed architecture reads the first k -columns of the dataset, and the GRF coding is applied to that input data (using one or more GRFPs) for obtaining the input firing times corresponding to those input data. Later, the architecture takes the input firing times and generates the network output (or output firing times). Depending of the layers implemented in the SNN, is the number of implemented SNLPs (at least one SNLP must be implemented for 2-layer SNN). This process is repeated until all the patterns in the dataset (or a number of patterns previously established) have been evaluated.

Gaussian Receptive Field Processor In figure 4, the main components of the GRFP are shown. The main components are described below:

- *Internal Control Unit*. This component receives control signal from Global Control Unit and generates the appropriate control signal for the components of the GRFP.
- *Maximum-Minimum-Range Computation Module (MMRCM)*. This module accesses to the dataset and computes maximum, minimum and range for each data column of the dataset. This module can be excluded from the architecture or not synthesized if these parameters are known or were pre-computed before coding. The results obtained by this module are stored in the Maximum-Minimum-Range Memory(MMRM).
- *Centroid Computation Module (CCM)*. This module computers the gaussian centroids as defined by equations 2 and 3. The total amount of centroids

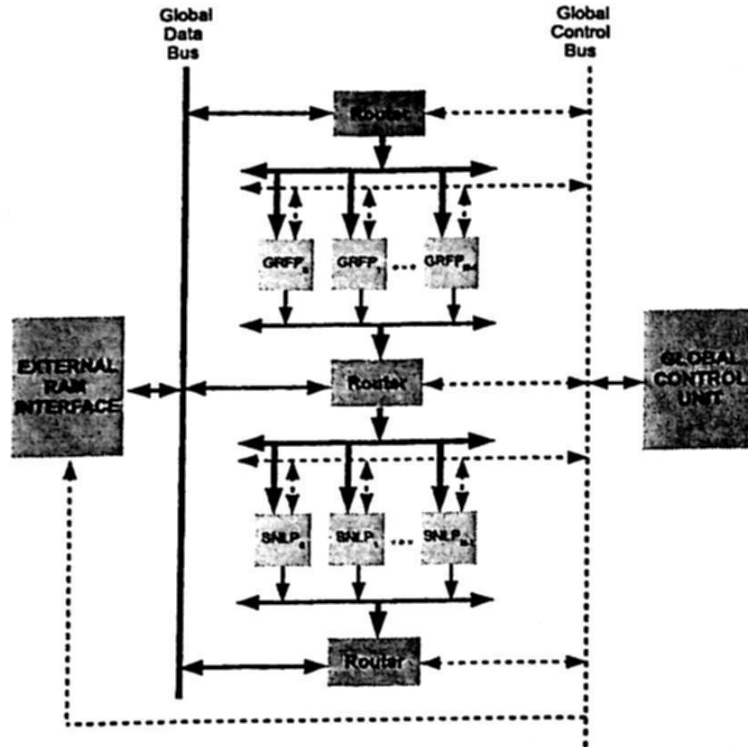


Fig. 3. Complete architecture with GRFPs and SNLPs modules

required depends of the number of gaussian fields required. The results obtained by this module are stored in the Centroid Memory (CM).

- *Parameters Memory (PM)*. The statistics and centroids are stored in this memory and later sent to each GP for the Gaussian Field Computation.
- *Temporal Registers (TR)*. Data from the input dataset are stored in an array of temporal registers. Each one of the GPs accesses to this register for obtaining the corresponding firing times according to the number of gaussian fields processor outputs.
- *Gaussian Processor (GP)*. This module performs the computation of each gaussian function of the gaussian array. Each GP reads data from all the TRs and obtains the Gaussian codification of that value, and later the results are stored in the corresponding FT.
- *Firing Times Memory (FTM)*. The firing times generated for each GP are stored in this memory. Later, the data are sent to the router for storing in external memory or to be sent to the SNLP.

The proposed architecture for the GRFP is designed to be flexible and modular depending of the required degree of parallelism. The parameters for architecture compilation that can be set for this processor are the number of gaussian fields, the width of the gaussian fields and the gaussian fields separation.

Spiking Neural Layer Processor (SNLP) This processor performs the recall phase in SNNs, which consists on the firing time computation of both hidden

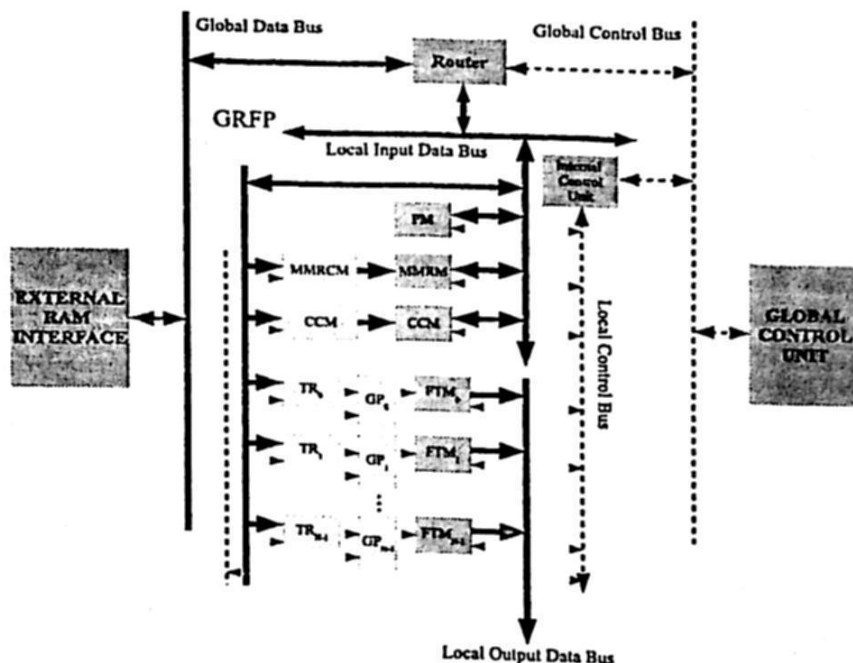


Fig. 4. Main components of a GRFP

and output neurons layers. The neural computation takes as input the firing times generated by the GRFPs. In the actual implementation, this processor is designed for working only with feed-forward SNNs, but it can be modified for supporting other connectivity schemes. The quality and precision of the GRF coding performed by the GRFPs is a critical factor for the network performance (related to the classification accuracy, not performance). The processor contains a set of modules called Neural Processors (NPs), which are grouped in layers depending of the requirements of the implemented SNN. A SNLP contains a set of fixed NPs, which are assigned for computing the neuron firing time for exactly one network layer. As reported in several applications, for the proposed architecture only two layers of Spiking neurons are defined (and only 2 SNLPs are implemented), but the proposed architecture is designed for implementing more neuron layers by defining more than one SNLP.

4 Implementation and Results

4.1 FPGA implementation

The proposed architecture was implemented and synthesized on an Alphasdata ADM-XPL board, which hosts a Virtex- II PRO FPGA. The hardware resources for the target FPGA device are shown in table 1. The target platform has a PCI interface, thus it is hosted on a desktop PC. The proposed architecture is fully modeled using the Handel-C Hardware Description Language.

The proposed architecture was designed for supporting as many processors as possible, limiting the implementation to the hardware resources available on the

target FPGA. In order to explore the parallelism tradeoffs, different implementations were synthesized. In table 2, the hardware resource utilization for each implementation is shown. For validating the GRF coding, only one GRFP module was synthesized (In table 2, second column). For validation the GRF coding and network performance, several NPs (4,8, and 16 NPs) per SNLP (only 2 SNLPs) were synthesized (In table 2 from third to fifth column).

Table 1. Characteristics of the target FPGA device

FPGA	4-Input LUTs	Slice-FFs	Total Slices	Total BRAMs	Total MULT18x18s
xc2vp30-6ff896	27,392	27,392	13,696	136	136

4.2 Performance results

The implemented architecture was tested using several variations of the input parameters. The input simulation parameters that can be set are:

- Number of examples (rows) and variables (columns) of the input dataset.
- Number of gaussian fields used for each variable (the number of input neurons is given by the product of total variables by the number of gaussian fields associated to each variable).

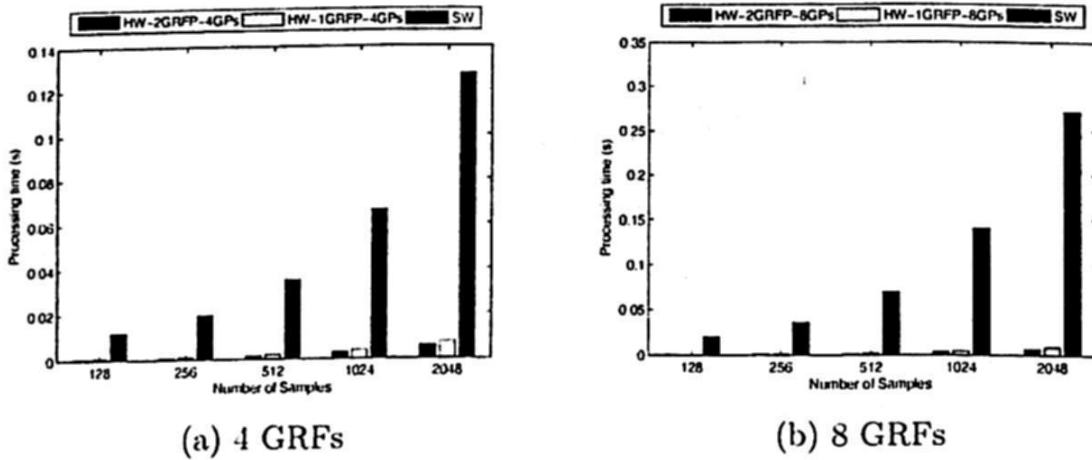


Fig. 5. HW versus SW execution time comparison for the GRFPs

Execution time comparison for the coding module is shown in figure 5, where the “SW” label means execution time in software, on a PC with a Pentium-4 processor running at 2.66 GHz. The “HW-2GRFP-4GPs” label means execution

time in hardware using 2 GRFPs with 4 GPs, the “HW-1GRFP-4GPs” label means execution time in hardware using 1 GRFP with 4 GPs, the “HW-2GRFP-8GPs” label means execution time in hardware using 2 GRFPs with 8 GPs and the “HW-1GRFP-8GPs” label means execution time in hardware using 1 GRFP with 8 GPs. The number of computed GRFs is 4 GRFs for figure 5(a) and 8 GRFs for figure 5(b).

Execution time comparison for the overall architecture is shown in figure 6, where the “SW” label means execution time in software, on a PC with a Pentium-4 processor running at 2.66 GHz. The HW-XP means execution time in hardware with 2 SNLPs with X NPs each. The experiments were performed with real data obtained from standard datasets used in machine learning applications. For the first experiment, the Iris dataset from the UCI Repository of Machine Learning Databases [12], with 150 instances and 4 discrete attributes was used. The performance for several topological variations of the implemented SNN for the Iris dataset is shown in figure 6(a). For the second experiment, the Wisconsin Breast Cancer (WBC) dataset from the UCI Repository of Machine Learning Databases [12], with 699 instances and 9 attributes was used. The performance for several topological variations of the implemented SNN for the WBC dataset is shown in figure 6(b). The reported execution time is the total amount of time required for both hardware and software implementations to perform the network output computation for the entire dataset.

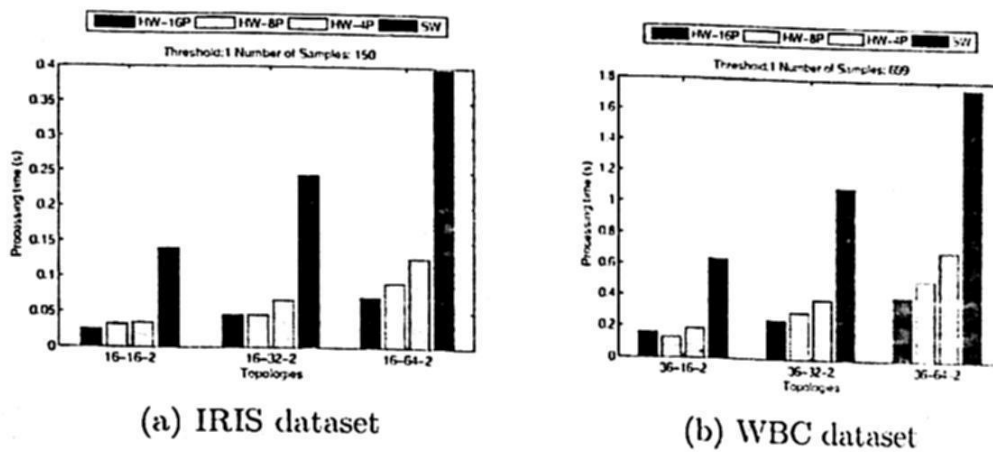


Fig. 6. HW versus SW execution time comparison (real datasets)

4.3 Discussion

The architecture described in this paper is an extension of work reported in the past. A first attempt for implementing an FPGA-based architecture for SpikeProp is reported in [13], using the original Spikeprop algorithm reported in [3]. The proposed architecture was validated by obtaining the same classification results as the original application. In a second attempt [14], several improvements

Table 2. Hardware resources for the complete architecture

Resources	GRFs ONLY	8 NPs	16 NPs	24 NPs
Slices	4,166 30%	6,843 50%	8,225 60%	9,711 71%
Max clock freq (Mhz)	66,8	66,84	65,17	64,22
Gate Count	458,161	2,523,393	4,155,493	5,791,707

to SpikeProp were integrated into the architecture previously proposed. It is specially interesting the reduction of the number of synaptic terminal from 16 synaptic connections in the original SpikeProp algorithm to 2 synaptic connections reported in [15], and the adaptations of RProp and QuickProp algorithm established in [4]. The recall phase is fully implemented, but the learning is validated only by an off-chip execution. In the first and second attempts, the coding problem was not explored. In a third attempt [16], a hardware architecture for the gaussian receptive fields coding is reported. Tests with randomly generated datasets are reported. In a fourth attempt [17], a generic hardware core for obtaining the network output for multilayer SNNs based on SpikeProp algorithm is reported. A previous version of the learning implementation is reported, as well as an estimation of hardware resources and performance. In this work, an integrated architecture and new improvements are reported. The novelty of the present work consists on the flexibility and scalability of the proposed architecture, as a compact core that can be used for a wide variety of classification problems, allowing the tuning of several parameters for both coding and recall phases of SNNs and improving in performance.

In the first part of the chain of processing of the proposed architecture, the GRF coding is performed using GRFPs. When synthesizing the proposed architecture with only GRFPs (the SNLPs were excluded), execution time results were obtained, and the obtained performance improvement falls in the range from 4X to 16X depending of the number of implemented GPs. The maximum number of GRFs that can be computed is established to 4, as proposed in [4], since 4 is good enough for obtaining an acceptable network performance, but if more GRFs are required, the architecture can be synthesized for fitting that requirement. These improvement rates are obtained because each GRF is mapped in one separated module, and several patterns (when adding more GRFPs) can be processed in parallel. About 30% of the target FPGA device is used when only GRF coding is implemented. The implemented GRFP uses only 4 GPs. If more GRFPs and GPs are required, then the architecture can be extended, but this involves more hardware resources for the GRFs coding. Using the implemented cores only for coding, at least a 2X performance improvement is obtained, and

for the performed experiments, a maximum performance improvement of 16X is obtained.

In the second part of the processing chain, the SNLP are added to the implemented architecture. The obtained performance improvement is from 4X to 20X depending of the number of NPs. About 50% of the FPGA device is used, with 1 GRFP and 2 SNLPs implemented. This core combination can be considered as the “minimal” implementation that functionally can achieve all the computations required for the SNNs described in this work. When using the “minimal” architecture, at least a 3.5X performance improvement is obtained for the smallest tested network, and a maximum 9.5X performance improvement for the large network tested in this work. When using more that 4 NPs for SNLP, the area required for the design is increase by a 1.5 factor. The limitation for the implementation of more SNLPs is the target FPGA available resources. The used hardware resources scale linearly with the number of neurons, and the performance decreases linearly but with a small slope compared with the hardware resources increasing. The performance improvement is very good (in the order of 7X to 9X), when using regular topologies (networks with the same number of neurons in each layer), and not too good (about 2X - 4X) when using an unbalanced number of neurons in the layers (see figures 6(a) and 6(b) for an example).

5 Conclusions and future work

A scalable and modular hardware core for SNNs has been proposed. The architecture is based in two phases: one preprocessing phase (GRF coding) and one processing phase for multilayer SNNs: the recall phase. For each one of these phases, a hardware core is proposed. For the coding phase, the GRFs is performed by a set of processors called GRFPs, a performance improvement of at least 4X is obtained. For the GRFs combined with the Recall Phase, a performance improvement of at least 3.5 X is obtained. The improvement to the first hardware core is reported in this work, while the second hardware core was left unmodified from the original version reported previously in [17]. The integration of both cores is fully documented in this work, as well as performance and resource utilization statistics. The proposed results shown an important improvement in performance with respect to SW-based implementation, and the modularization of the proposed architecture allows to implement the proposed architecture using larger FPGA devices.

As future work, the integration of both learning and recall blocks with the implemented modules is proposed. The reuse of hardware resources can improve the execution time, allowing to have a better network performance. If all the stages of the neural processing are implemented on-chip, the compactness of this core can allow the implementation of high-performance classifying systems. The testing of the proposed architecture with more standard datasets used in Machine Learning algorithms is required, and the integration of the core with more challenging neural networks applications, like speech recognition is needed.

References

1. Maass, W.: Networks of spiking neurons: the third generation of neural network models. *Transactions of the Society for Computer Simulation International* 14(4) (1997) 1659–1671
2. Gerstner, W., Kistler, W.: *Spiking Neuron Models: An Introduction*. Cambridge University Press, New York, NY, USA (2002)
3. Bohte, S.M., Kok, J.N., Poutre, H.L.: Spikeprop: Error-backpropagation for in multi-layer networks of spiking neurons. *Neurocomputing* 1–4(48) (November 2002) 17–37
4. S. McKennoch, D.L., Bushnell, L.G.: Fast modifications of the spikeprop algorithm. *IEEE World Congress on Computational Intelligence (WCCI)* (July 2006)
5. Wu, Q.X., McGinnity, T.M., Maguire, L.P., Glackin, B.P., Belatreche, A.: Learning under weight constraints in networks of temporal encoding spiking neurons. *Neurocomputing* 69(16–18) (2006) 1912–1922
6. Moore, S.C.: *Back-Propagation in Spiking Neural Networks*. Master's thesis, University of Bath, United Kingdom (2002)
7. Benjamin, S., Jan, V.C.: Backpropagation for population-temporal coded spiking neural networks. In: *Proceedings of the 2006 International Joint Conference on Neural Networks, IEEE* (1 2006) 3463–3470
8. Gerstner, W.: *Populations of spiking neurons*. Maass, W. and Bishop, C., editors, MIT-Press, Cambridge (1999)
9. Johnston, S., Prasad, G., Maguire, L.P., McGinnity, T.M.: Comparative investigation into classical and spiking neuron implementations on fpgas. In: *ICANN* (1). (2005) 269–274
10. Maass, W., Bishop, C.M., eds.: *Pulsed neural networks*. MIT Press, Cambridge, MA, USA (1999)
11. Jahnke, A., Schönauer, T., Roth, U., Mohraz, K., Klar, H.: Simulation of spiking neural networks on different hardware platforms. In: *ICANN '97: Proceedings of the 7th International Conference on Artificial Neural Networks, London, UK, Springer-Verlag* (1997) 1187–1192
12. Asuncion, A., Newman, D.: *UCI machine learning repository* (2007)
13. Nuño-Maganda, M., Arias-Estrada, M., Torres-Huitzil, C.: An efficient scalable parallel hardware architecture for multilayer spiking neural networks. In: *SPLCONF 2007*. (2007) 167–170
14. Nuño-Maganda, M., Arias-Estrada, M., Torres-Huitzil, C.: High performance hardware implementation of spikeprop learning: Potential and tradeoffs. In: *ICFPT07, Kytakyushu, Japan* (2007) 129–136
15. Benjamin, S., Jan, V.C.: Extending spikeprop. In Campenhout, J.V., ed.: *Proceedings of the International Joint Conference on Neural Networks, Budapest* (7 2004) 471–476
16. Nuño-Maganda, M., Arias-Estrada, M., Torres-Huitzil, C., Girau, B.: A population coding hardware architecture for spiking neural network applications. In: *SPLCONF 2009, São Carlos, Brazil* (2009) 83–87
17. Nuño-Maganda, M., Arias-Estrada, M., Torres-Huitzil, C., Girau, B.: Hardware implementation of spiking neural network classifiers based on backpropagation-based learning algorithms. In: *IJCNN09, Atlanta, U. S.* (2009) 2294–2301

An Intelligent Virtual Agent for Collaborative Learning looking to be part of the Team

Raúl A. Aguilar¹, Angélica de Antonio², Ricardo Imbert²
and Adriana Peña³

¹ Universidad Autónoma de Yucatán, Mathematics School
Periférico Norte Tablaje 13615, A.P. 172, Cordemex,
C.P. 97110, Mérida, México
avera@uady.mx

² Universidad Politécnica de Madrid, Computer Science School
Campus Montegancedo, 28660, Boadilla del Monte, Madrid, Spain
{angelica, rimberty}@fi.upm.es

³ 3DEVICE, S.A. de C.V.
Serapio Rendón 245, Centro, C.P. 58000, Morelia, México

Abstract. The Team Role Theory sustains that a balance among the members' preferences towards certain roles improves the team performance. In this paper an Intelligent Virtual Agent originally designed applying an emotion-based architecture to generate particular behaviors, is modeled according to compatible team role characteristics, in order to substitute a student when required, during the execution of a socio-technical task. An exploratory study is also briefly commented.

Keywords: Collaborative Virtual Environment, Computer Supported Collaborative Learning, Intelligent Virtual Agent.

1 Introduction

The reciprocal interplay between learning and working has contributed to the understanding to one another. In this context, the Team Role Theory although originally proposed for working teams, has being found beneficial for the construction of knowledge in Computer Supported Collaborative Learning (CSCL) environments [1].

While the main purpose of CSCL is to scaffold students in learning together effectively; its aim of research comprises not only the techniques but also the social organization to better support collaborative learning [2].

Some learning groups seem to interact naturally, even though, others struggle to maintain a balance of participation, leadership, understanding, and encouragement [3]. According to the Team Role Theory people tend to behave, contribute and interrelate with others at work in certain distinctive ways.

© M. Martínez, A. Alarcón (Eds.)
Advances in Computer Science and Engineering.
Research Computing Science 45, 2010, pp. 157-167

Received 30/04/10
Final version 19/05/10
Accepted 14/05/10

A balanced combination on these preferred ways of interacting with others while working is expected to result in a more successful team than those with an unbalanced composition [4].

On the other hand, Collaborative Virtual Environments (CVE) are a powerful tool for learning that combine virtual worlds in a distributed system, offering a shared space where the students can navigate, and interact with peers, data and objects through visual and auditory channels [5]. In Virtual Environments (VE), materials do not break or wear out; while they allow safe experiences of distant or dangerous locations and processes [6]. Their main uses are likely to be where spatial tasks are involved because they are commonly and predominantly visual; where co-presence is required; and where it is more effective or more enjoyable to carry out a task or activity in virtual than in real, for reasons of cost, safety or interpersonal difficulty [7].

CVE's characteristics make them proper for socio technical tasks (like training in coordinated situation such as rescue operations or enterprise logistic). While this type of tasks is accomplished on line, creating the plan or evaluating results could be carry out by an asynchronous participation, but when it comes to its execution, all the students may need to interact at the same time. In which case a pedagogical agent could take a student's place while at the same time plays the required team role for balance, an agent that will act accordingly to the requirements of the task and to be compatible with the other students' personal characteristics. The agent can either balance the team roles or replace a student.

2 Intelligent Virtual Agent Architecture for Team Roles

PANCHO (Pedagogical AgeNt to support Collaborative Human grOups) is an Intelligent Virtual Agent designed applying an emotion-based Architecture to generate particular behaviors [8], in this case, related to the team roles as defined by Belbin .

The Belbin's team roles [4, 9] is one of the more well known instruments related to personality and team preferences, the earliest and still the most popular categorization. Belbin proposed nine team roles classified in three types: People oriented (P), Cerebral (C), and Action oriented (A), see Table 1. A balanced team should have members of different kind of roles.

In order to identify the role a person will have in a team, Belbin [9] designed a questionnaire for the team member called the SPI: Self-Perception Inventory complemented by the group member, and confirmed by other questionnaire filled out by his/her peers.

Belbin [4] identified that certain roles are compatible or not with the others, and associated this compatibility to a hierarchical situation such as supervisor or peers. Because in a proper collaborative learning situation it is expected to avoid the hierarchical organization [10, 11], only the relation between the team roles in a flat structure are here presented in Table 2.

The team roles are closely related to people personalities. One of the personality models more used as reference is the OCEAN or Big Five Model [12], this model distills the differences between individual personalities into five basic personality factors very briefly explained: *extraversion*, socially compromised; *agreeable*, concerned with cooperation and social harmony; *conscientiousness*, impulses control; *neuroticism*, emotionally reactive; and *openness*, conventional.

Table 1. Belbin's Team Roles

Kind		Role	Description
(A)ction oriented	(SH)	Sharper	Dynamic, challenging. Has drive and courage to overcome obstacles.
	(IM)	Implementer	Disciplines, reliable, conservative. Turns ideas into practical actions.
	(CF)	Completer Finisher	Painstaking, conscientious, anxious. Searches out errors and omissions, delivers on time.
(P)eople oriented	(CO)	Coordinator	Mature, confident, a good chairperson. Clarifies goals, promotes decision making.
	(TW)	Team Worker	Cooperative, mild, perceptive, diplomatic. Listens, builds, averts friction.
	(RI)	Resource Investigator	Extrovert, enthusiastic. Explores opportunities. Develops contacts.
(C)erebral	(PL)	Plant	Creative, imaginative, unorthodox. Solves difficult problems.
	(ME)	Monitor-Evaluator	Sober, strategic, discerning. Sees all options.
	(SP)	Specialist	Single-minded, self starting. Dedicated. Provides Knowledge and skills in rare supply.

Lindgren [13] reported a rational analysis of those personality factors that influence the most each Belbin's Team Roles. Lindgren gave positive or negative weight to each personality factor for each team role as shown in Table 3.

Table 2. Team Roles relations in a flat structure organization

Role	Relation	
	Compatible with (Preferences)	Incompatible with
Sharper (SH)	Resource Investigator (RI)	Coordinator (CO) Team Worker (TW)
Implementer (IM)	Coordinator (CO) Resource Investigator (RI) Monitor-Evaluator (ME) Specialist (SP) Completer Finisher (CF)	Implementer (IM) Plant (PL)
Completer Finisher (CF)	Implementer (IM)	Resource Investigator (RI) Monitor-Evaluator (ME)
Coordinator (CO)	Team Worker (TW) Implementer (IM)	Sharper (SH)
Team Worker (TW)	Team Worker (TW) Plant (PL)	Sharper (SH)
Resource Investigator (RI)	Team Worker (TW) Implementer (IM)	Completer Finisher (CF) Specialist (SP)
Plant (PL)	Coordinator (S) Resource Investigator (RI) Team Worker (TW)	Monitor-Evaluator (ME) Plant (PL) Specialist (SP) Implementer (IM)
Monitor-Evaluator (ME)	Coordinator (S) Implementer (IM)	Completer Finisher (CF) Monitor-Evaluator (ME) Plant (PL)
Specialist (SP)	Implementer (IM) Team Worker (TW)	Plant (PL) Resource Investigator (RI)

Table 3. Lingren's [13] relation between Team Roles and Personality Factor

Team Role	Personality Factor				
	Extraversion (I)	Agreeableness (II)	Conscientiousness (III)	Neuroticism (IV)	Openness (V)
Sharper (SH)	9	-5	1	-1	0
Implementer (IM)	0	0	7	0	-3
Completer Finisher (CF)	0	1	8	-2	1
Plant (PL)	0	0	-2	-1	11
Monitor- Evaluator (ME)	-3	-4	2	2	3
Specialist (SP)	1	-2	8	0	1
Coordinator (CO)	3	3	2	4	2
Resource Investigator (RI)	8	0	-3	1	2
Completer Finisher (CF)	-2	5	0	-1	2

Accordingly, the virtual agent *Pancho* defining characteristics $DC(Pancho)$ as part of the team will be: its kind of team role $KR(Pancho)$; its team role $TR(Pancho)$; and the personality factors that correspond to them $P(Pancho)$, that is: $DC(Pancho) = KR(Pancho) \quad TR(Pancho) \quad P(Pancho)$.

We will first establish the agent's kind of role that will balance the team, then the specific team role according to the defined kind, and to be compatible with the other team members, and finally the personality factors that match the team role and will determine its socio interaction.

2.1 The Team Roles Balance

A small group is recommended for collaborative learning in order to give all of its members the opportunity to participate. There is not an exact specification about how many members a group should have in order to be called a small, but in groups with more than five members there is a general complain about participation restrictions. [14]. A group integrated by three members was selected to exemplified the approach; according to Bean [15] an optimal group size for the workgroup. To balance the team, the first defining characteristic is the kind of role (KRole) that the agent must play, considering T_x, T_y Human Group with $x \neq y$:

$RealTeam = \{KRole(Tx), KRole(Ty)\}$

$IdealTeam = \{A, P, C\}$

If $KRole(Tx) \neq KRole(Ty) \neq KRole(Pancho) \neq (IdealTeam - RealTeam)$.

If $KRole(Tx) = KRole(Ty) \neq KRole(Pancho) = p \neq KChoice[p] > KChoice[q]$
where $p \neq q, p \neq KRole(Tx)$ and $q \neq KRole(Tx)$.

Then by an analysis of incompatibilities the agent team role might assume is, for example, if the agent got a People Oriented team role then:

If $(KRole(Pancho) = P) \neq (TR(Tx) = SH \neq TR(Ty) = SH) \neq TR(Pancho) = RI$.

If $(KRole(Pancho) = P) \neq (TR(Tx) \neq \{CF, SP\} \neq TR(Ty) \neq \{CF, SP\}) \neq TR(Pancho) \neq \{CO, TW\}$

If $(KRole(Pancho) = P) \neq (TR(Tx) \neq \{SH, CF, SP\}) \neq (TR(Ty) \neq \{SH, CF, SP\}) \neq TR(Pancho) \neq \{CO, RI, TW\}$

Once the agent team role is defined then its personal factors can be shaped.

2.2 Shaping the Agent's Personality

People tendency does not necessarily mean they will always act in the same way, although they may have a deep marked tendency. The heuristic adopted for generic purposes was to restrict the values for each of the five factors as presented in Table 3, allowing the tendency to move accordingly to a range of possibilities. Through fuzzy logic linguistic labels, the quantitative of the personality factors are transformed to linguistic variables (as in [16]), see Figure 1. The qualitative linguistic values that each personality factor will take accordingly to the Lingren's weights (Table 3) are presented in Table 4.

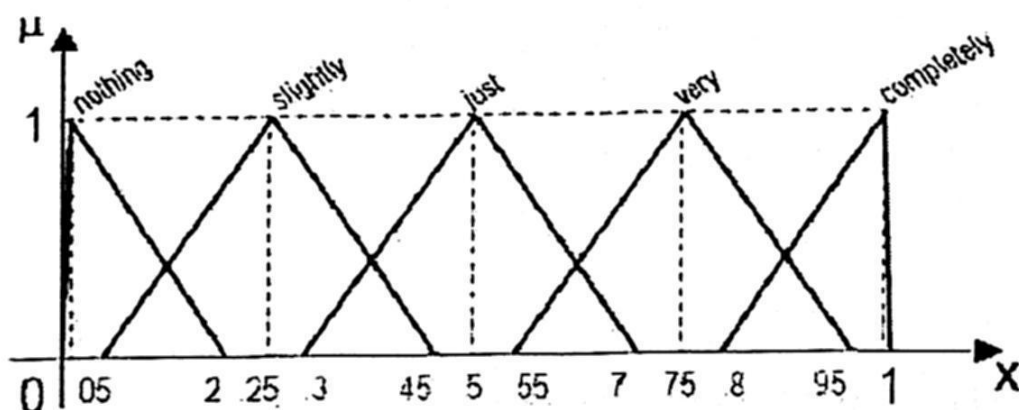


Fig. 1. Semantic Model for the factor degrees

Table 4. Personality factors heuristic in a qualitative domain

Quantitative Domain	Heuristic	Qualitative Domain				
		N	S	J	V	C
[0]	Without influence. All values	✓	✓	✓	✓	✓
[-2, -1]	Minimum influence. Four values according to the influenced pole	✓	✓	✓	✓	
[1, 2]			✓	✓	✓	✓
[-4, -3]	Certain influence. Three values according to the influenced pole	✓	✓	✓		
[3, 4]				✓	✓	✓
[-7, -6, -5]	Significant influence. Two values according to the influenced pole	✓	✓			
[5, 6, 7]					✓	✓
[-11, -10 -9, -8]	High influence. One extreme value	✓				
[8, 9, 10, 11]						✓

The range for the restricted values for the personality factors of each team role for the People oriented, Cerebral, and Action oriented types are shown in Table 5.

The *coordinator* differs from other People oriented roles for his/her emotional stability; he/she is a mature and confident person with high values for all the personality factors. The *resource investigator* is enthusiastic and communicative person (extraversion: completely) that develops contacts. The third People oriented role, the *team worker* is cooperative, mild and diplomatic (agreeableness: very, completely) but he/she averts friction which may provoke edginess (neuroticism: just).

The cerebral type of role is creative, imaginative, and unorthodox (openness: completely) but he/she might be careless with the practical tasks. One main *specialist* characteristic is that he/she is highly dedicated (conscientiousness: completely). As of the *monitor-evaluator*, he/she is strategic and a discerning person, although not very sensible with his/her peers' problems (extraversion and agreeableness: nothing, slightly, just).

Table 5. Personality factors related to People oriented, Cerebral and Action oriented roles

Personality Factors	People oriented roles		
	Coordinator	Resource Investigator	Team Worker
Extraversion	[J, V, C]	[C]	[N, S, J, V]
Agreeableness	[J, V, C]	[N, S, J, V, C]	[V, C]
Conscientiousness	[S, J, V, C]	[N, S, J]	[N, S, J, V, C]
Neuroticism	[J, V, C]	[S, J, V, C]	[N, S, J, V]
Openness	[J, V, C]	[S, J, V, C]	[S, J, V, C]
Personality Factors	Cerebral roles		
	Plant	Monitor-Evaluator	Specialist
Extraversion	[N, S, J, V, C]	[N, S, J]	[S, J, V, C]
Agreeableness	[N, S, J, V, C]	[N, S, J]	[N, S, J, V]
Conscientiousness	[N, S, J, V]	[S, J, V, C]	[C]
Neuroticism	[N, S, J, V]	[S, J, V, C]	[N, S, J, V, C]
Openness	[C]	[J, V, C]	[S, J, V, C]
Personality Factors	Action oriented roles		
	Shaper	Implementer	Completer Finisher
Extraversion	[C]	[N, S, J, V, C]	[N, S, J, V, C]
Agreeableness	[N, S]	[N, S, J, V, C]	[S, J, V, C]
Conscientiousness	[S, J, V, C]	[V, C]	[C]
Neuroticism	[N, S, J, V]	[N, S, J, V, C]	[N, S, J, V]
Openness	[N, S, J, V, C]	[N, S, J]	[S, J, V, C]

In the group of Action oriented type, the *shaper* is clearly sociable when trying to achieve a goal (extraversion: completely), but his/her cold and critical personality can create antipathy. The *completer finisher* and *implementer* have personality factors similarities, both are responsible and not clearly sociable, but while *implementers* do

not like changes (openness: nothing, slightly, just), the *completer finishers* are conscientious and perfectionist (conscientiousness: completely).

3 ¿Is Pancho my Teammate, a Coordinator or an Implementer?

The agent was implemented in an experimental CVE with ludic characteristics, which goal is to transport a diplomatic by plain, through an enemy zone constantly monitored by other plains and ships with a determined path.

In the CVE, the agent performs the team leader either as a Coordinator or as an Implementer. These roles were selected because they are different kind, the coordinator is People oriented while the Implementer is Action oriented. Belbin [4] described the Coordinator personality as one likely to be a more natural team leader. While the Implementer best team quality is to execute the planned actions, and his/her leadership may lack of spontaneity. Their personality factors were settled as shown in Table 6.

Table 6. Personality factors applied for the Coordinator and Implementer roles

Personality Factors	Coordinator	Implementer
Extraversion	Very	Just
Agreeableness	Very	Just
Conscientiousness	Just	Completely
Neuroticism	Very	Just
Openness	Very	Nothing

As mentioned, the intelligent virtual agent knowledge about the task at hand primarily guides its course of action. And, it is part of an emotional-based architecture, this means it reacts not only accordingly to its personality but also to emotions such as joy, trust, or fear; attitudes towards its peers; and its physical state like being tired or thirsty (see for further details [16]). However, the focus here is its team role characteristics. The actions that the agent takes according to its team role are presented in Table 7.

Table 7. The agent's actions taken accordingly to the team role

Situation	Coordinator action	Implementer action
dangerous type	Takes an immediate action (even if it is not the best one). Sends a message trying to release the team tension	Delays to take a decision until it gets a clear definition of the dangerous situation
risky type	Elaboration of a number of possible actions	Elaboration of an efficient combination of concrete actions
a determined reached percentage of the goal	Sends messages informing the advances to the team's goals	None

With the idea to understand how the human members of the team perceived the Pancho's team role, a first exploratory study, briefly commented in the next section, was conducted.

3.1 Exploratory Study

Four teams with two compatible different roles were formed with Computer Science students. Each team completed the task with Pancho as the third member of the team twice, each one with Pancho in a different role (Coordinator and Implementer).

The Belbin's [9] SPI: Self-Perception Inventory was adapted for three roles: the coordinator, the implementer and the plant. And the participants qualified with it the agent's profile.

Results are shown in Table 8. The students identified correctly the Pancho's team role as Coordinator only two of the eight times (25% of the times), as Implementer half of the times (four of eight times) and as plant 25% of the time. The Implementer was identified correctly half of the time, twice it received the same qualification as Implementer or Coordinator, and one the same for the three roles.

The results show a tendency to identify the agent as an Implementer regardless to its defined characteristics. The one condition that clearly affected the results is the small size of the sample, but other conditions that may present more accurate results are a larger number of different actions the agent could take, or a greater number of roles that the agent could perform.

Table 8. The participants qualifications for the agent's team role

Student	Coordinator (CO)Team Role			Implementer (IM)Team Role		
	CO	IM	PL	CO	IM	PL
01	9	11	10	11	12	7
02	9	10	11	10	10	7
03	7	19	4	12	12	10
04	7	12	11	11	12	6
05	6	10	14	9	15	7
06	9	12	9	11	12	6
07	18	8	4	19	10	7
08	18	7	5	13	13	4

4 Discussion and Ongoing Work

In this paper we presented the rationalization for modeling an intelligent virtual agent, accordingly to a certain team role, with the intention to be compatible with a group of students that take care of a socio-technical task and during its execution. The task is meant to be carried out in a CVE for learning. The agent can take the place of a student in order to complement the group for training purposes or to balance the team.

The implementation of the approach was made in a CVE giving to the agent two types of team role. An initial exploratory study was conducted, although with no posi-

tive or certain results its main outcome is the insights that will help to create a more adequate experimental design.

Currently we are working on the design of an experiment that can answer questions about the improvement in the team performance by using our agent.

References

1. Roberts, A. G.: Team Role Balance: Investigating Knowledge-Building in a CSCL Environment. Unpublished Ph.D. Thesis. (2007)
2. Hsiao, L. J.: CSCL Teories. University of Texas at Austin, Web Page <http://www.edb.utexas.edu/csclstudent/Dhsiao/theories.html#top> (2005)
3. Soller, A.: Supporting Social Interaction in an Intelligent Collaborative Learning System. *International Journal of Artificial Intelligence in Education*, **12** (2001) 40-62
4. Belbin, M.: Team Roles at Work. Oxford, Elsevier Butterworth Heinemann. (1993)
5. Churchill, E. F., & Snowdon, D.: Collaborative Virtual Environments: An Introductory Review of Issues and Systems. *Virtual Reality: Research, Development and Applications*, **3** (1998) 3-15
6. Bricken, M.: Virtual Reality Learning Environments: Potentials and Challenges. *Computer Graphics*, **25** (1991) 178-184
7. Spante, M., Heldal, I., Steed, A. et al.: Strangers and Friends in Networked Immersive Environments: Virtual Spaces for Future Living. (2003)
8. Aguilar, R.A., de Antonio, A., Imbert, R.: PANCHO needs Models of Collaborative Human Groups: A Mechanism for Teams Modeling. *Research in Computing Science*, **34**, (2008) 299-310.
9. Belbin, M.: Management Teams. John Wiley & Sons, New York. (1981)
10. Collazos, C. A., Guerrero, L. A., Pino, J. A. et al.: Evaluating Collaborative Learning Processes using System-Based Measurement. *Educational Technology & Society*, **10** (2007) 257-274
11. Dillenbourg, P.: What do you mean by collaborative learning?. In P. Dillenbourg (Ed) *Collaborative-learning: Cognitive and Computational Approaches*. Oxford: Elsevier (1999)
12. Digman, J. M.: Emergence of the Five-Factor Model. *Annual Review of Psychology*, **41** (1990) 417-440
13. Lindgren, R.: Meredith belbin's team roles viewed from the perspective of the big 5 : A content validation. *Universitetet i Oslo, Oslo* (1997)
14. Napier, R., & Gershenfeld, M.: Groups: Theory and experience. Houghton Mifflin, Boston (1975)
15. Bean, J. C.: Engaging ideas: The profesor's guide to integrating writing, critical thinking, and active learning in the classroom. Jossey-Bass, San Francisco (1996)
16. Aguilar, R. A., de Antonio, A., Imbert, R.: Searching Pancho's Soul: An Intelligent Virtual Agent for Human Teams. *Proceedings of the Electronics, Robotics, and Automotive Mechanics Conference* (2007) 568-571

Distributed System for Assessment of Water Quality in Shrimp Aquaculture Systems

José Juan Carbajal Hernández¹, Raúl A. Valero Cruz²
and Mauricio Suárez López²

¹Department of Real Time Systems and Modeling, Centre of Computer Research – IPN
Av. Juan de Dios Bátiz, Col. Nva. Industrial Vallejo, México D.F., México.

²Departamento de Ingenierías, Universidad del Valle de México
Paseo de las Aves No. 1, San Mateo Nopala, C.P. 53220,
Naucalpan de Juárez, Edo de México.
juancarvajal@sagitario.cic.ipn.mx

Abstract. Water quality in aquaculture systems must be under control, a disestablished ecosystem can be harmful for organisms. This work presents a new tool for assessment of the ecosystem status based on a distributed system, whose was developed in three phases: measurement (sensor), data acquisition (conditioning and analog to digital converter), and signal processing (software). A fuzzy inference System processes environmental information using a reasoning process. Potential negative situations and harmful combinations between physical-chemical variables are detected, providing a final water quality index, which describes in a status level of the ecosystem (excellent, good, regular and bad). A user interface was build for an easily water management information of the assessed ponds.

Keywords: Water quality, fuzzy inference, aquaculture, artificial intelligence.

1 Introduction

The water management is an important factor in shrimp aquaculture where the ecosystem must be under control. A disestablished habitat is not conducive for a good farming, also an organism with a weakened immunological system is more likely for getting Sick [1](for example Taura virus, Mancha Blanca, Cabeza Amarilla, Etc.).

The environmental variables have some concentration limits, where low or high concentrations (depending of the variable) can be harmful for the organism [2], [3], [4]. Following this behaviors, it is possible to implement a model in the attention that those limits and changes in the variables can be used for determining when a concentration is good or bad for shrimp, and how the combination of the variables affects the water quality in the artificial shrimp habitat. This strategy will decrease the negative situations; consequently also it will decrease the stress in the organism, and low mortality rates.

This research is based on analyze the water quality of *Litopenaeus vanammei*

shrimp, whose is cultivated in farms located in Sonora, Mexico, therefore toxic concentrations for this organism will be analyzed for the construction of the distributed system.

2 Distributed system

The production system and methods for Central America monitors in shrimp aquaculture have different frequencies of environmental variables. Dissolved oxygen, temperature and salinity are monitored daily; pH, ammonia, nitrate, turbidity and algae counts are measured weekly. In this work the pH variable is measured daily in order to control possible ammonia concentrations. Chemical analyses do not come into consideration for water quality management on a routine bases [5]. In attention of those variables with a higher frequency of measuring, the distributed system is implemented using this set.

The monitoring of physical-chemical variables can be classified in three phases: measurement (sensor), data acquisition (conditioning and ADC), and signal processing (software), Fig. 1 shows this process.

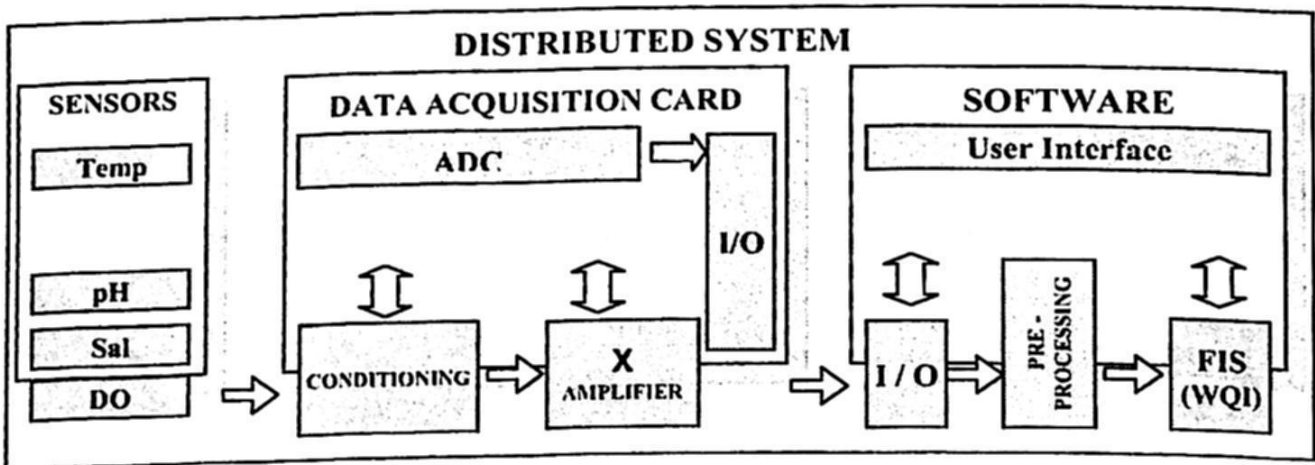


Fig. 1. Architecture of the distributed system for assessment of water quality.

3 Sensors and data acquisition

Different environmental sensors was used for measuring the physical-chemical variables, an explanation of how they were coupled in the distributed system is presented in this section.

pH sensor

The sensor used for measuring the pH was a Signet 2717, which is characterized for having a bulb ORP electrode with bulb protection and a preamplifier [6], whose has a proportionality expressed as follows:

$$pH = (-59)[Lecture(mV)] + 413 \quad (1)$$

The conditioning is a treatment of the original signal for a better lecture of the response. The pH conditioning consist on amplify the mV output in factor of $\times 10$ (Fig. 2), where an INA118 low noise instrumentation amplifier was used.

Temperature sensor

For monitoring temperature a LM35 sensor was used, whose output voltage is linearly proportional to the Celsius (Centigrade) temperature. The measurement range is between $-55 - 150^\circ\text{C}$, with a linear response of $+10.0 \text{ mV}/^\circ\text{C}$, as follows:

$$^\circ\text{C} = Lecture(mV) \times 100 \quad (2)$$

The temperature conditioning consists on amplifying the mV output in factor of $\times 10$ (Fig. 2), as the pH an INA118 low noise instrumentation amplifier was used.

Salinity sensor

Salinity was measured using a conductivity sensor Signet 2819, which is characterized for presenting a high resistance in a range between $18.2 \text{ M}\Omega$ to $10 \text{ K}\Omega$ for salinity between 0.02 to 50 mg/L [7]. Conductivity sensor conditioning consist on adapt a voltage divisor, measuring indirectly de resistance with a voltage produced by the sensor Fig. 2.

$$^\circ\text{C} = (-2.7378 \times 10^{-10})[Lecture(\Omega)] + 5 \times 10^{-3} \quad (3)$$

Dissolved Oxygen sensor

Dissolved oxygen was measured using a TruDO sensor, which is used for the measurement of the amount of dissolved oxygen present in a unit volume of water. DO sensors do not measure the actual amount of oxygen in water, but instead measure the partial pressure of oxygen in water, which is dependent on both salinity and temperature. The sensor output is a current that can be calculated from:

$$i_d = 4 \frac{A \times F \times P_{O_2} \times P_m(t)}{d} \quad [\text{Amp}] \quad (4)$$

where Faraday's constant, F , is $9.64 \times 10^4 \text{ [C/mol]}$, $P_m(t)$ is the permeability of the membrane (which is a function of temperature), A is the surface area of the noble metal electrode, P_{O_2} is the partial pressure of oxygen, and d is the thickness of the membrane [8]. Dissolved oxygen conditioning consists on convert current to voltage using a LM308 operational amplifier (Fig. 2).

Data acquisition

Data acquisition from sensors must be transmitted to the PC, this process can be implemented using a microcontroller PIC18F2550, and whose has an USB interface that can be programmable using the MPLAB environment (MPLAB, 2009) [9]. The circuit of the USB interface is shown in Fig. 2.

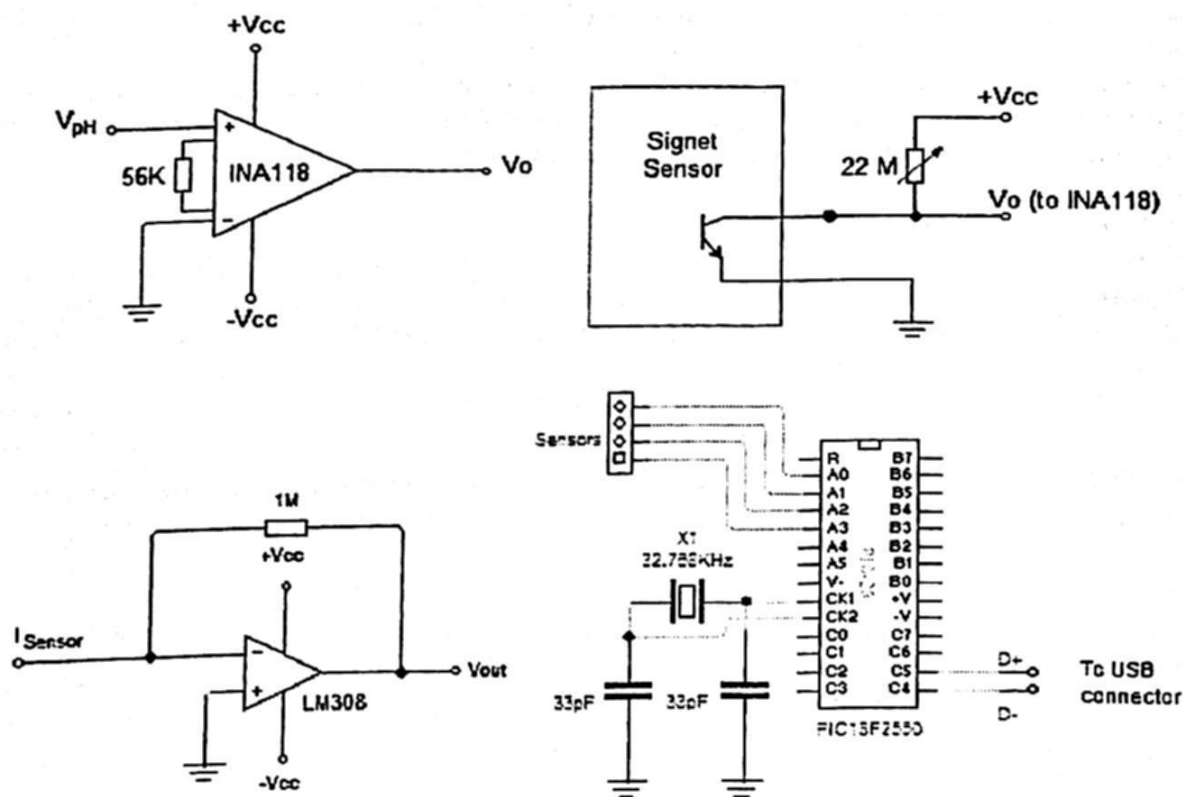


Fig. 2. Conditioning diagrams for sensors; a) low noise instrumentation amplifier (x10); b) Conductivity sensor connection, c) current to voltage amplifier (x1000) for dissolved oxygen, d) a PIC18F2550 is used as interface to ADC conversion and USB communication.

4 Signal Processing

4.1 Physical – chemical classification

In order to classify the behavior of a physical-chemical variable it is needed to define the ranges of optimal or harmful concentrations. The classification levels of the physical-chemical variables (status) are defined in Table 1.

Table 1. Classification levels, tolerances and limits of physical-chemical variables.

Variables	Hypoxia/Acid	Low	Normal	High	Alkaline	Tol	Lím
Temp (°C)	-----	0 – 23	23 - 30	30 - ∞	-----	±1	±1
Sal (mg/L)	-----	0 – 15	15 - 25	25 - ∞	-----	±1	±1
DO (mg/L)	0 - 3	3 – 6	6 - 10	10 - ∞	-----	±0.5	±0.5
PH	0 - 4	4 – 7	7 - 9	9 - 10	10 - 14	±0.5	±0.5

4.2 Fuzzy Inference Systems (FIS)

The Fuzzy inference systems (FIS) theory was applied in this study providing a non-linear relationship between input sets (Physical-chemical variables) and output set (Water Quality Index) [10], [11]. A FIS works in three phases: first it transforms real values in fuzzy outputs for the system using membership functions. Second phase process the fuzzy outputs using a reasoning process based on rules. Finally the output rules are aggregated to create a final output, which is used to determine a [0, 1] index; A Mamdani fuzzy inference system was developed in this work (Fig. 3).

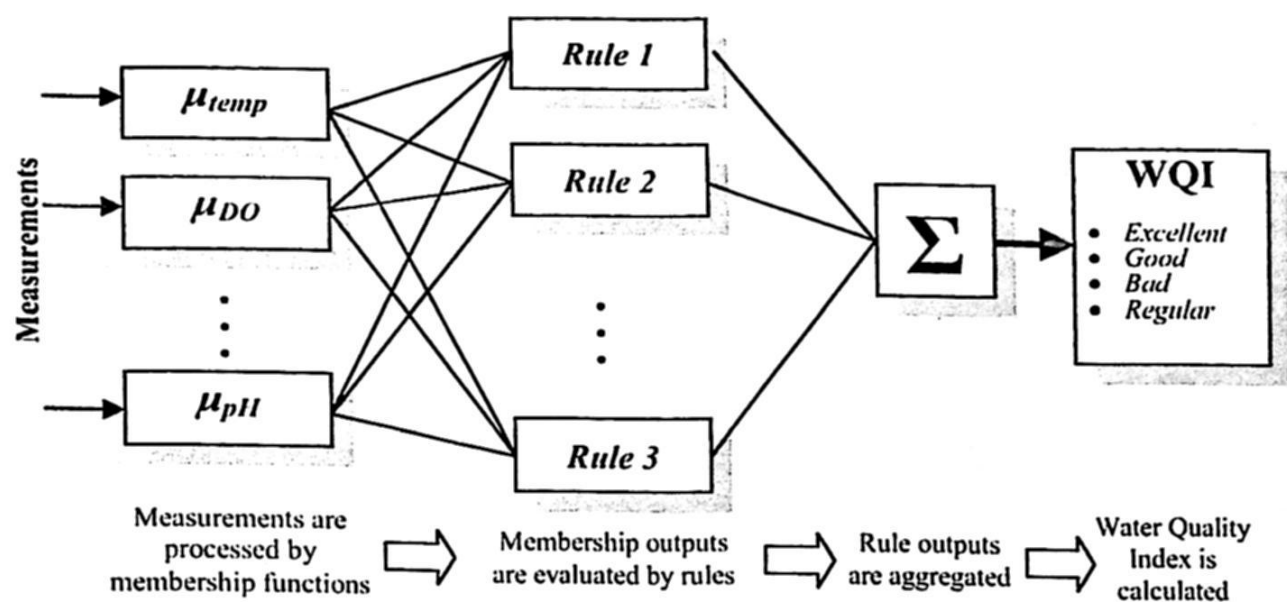


Fig. 3. Architecture of the Fuzzy Inference System applied to the water quality problem in shrimp farms.

Membership functions

Membership functions transform real measurements in [0, 1] indices, those can be implemented in different ways [12]. Expressions of fuzzy memberships are implemented as trapezoidal functions and they can be mathematically expressed as:

$$\mu_{WQI}(x, a, b, c, d) = \min \left\{ \frac{x - a}{b - a}, 1, \frac{d - x}{d - c} \right\} \quad (5)$$

where a , b , c and d are the membership parameters, x is the evaluated variable. Ranges, limits and tolerances in Table 1 are used to build the trapezoidal membership functions, whose are showed in Fig. 4.

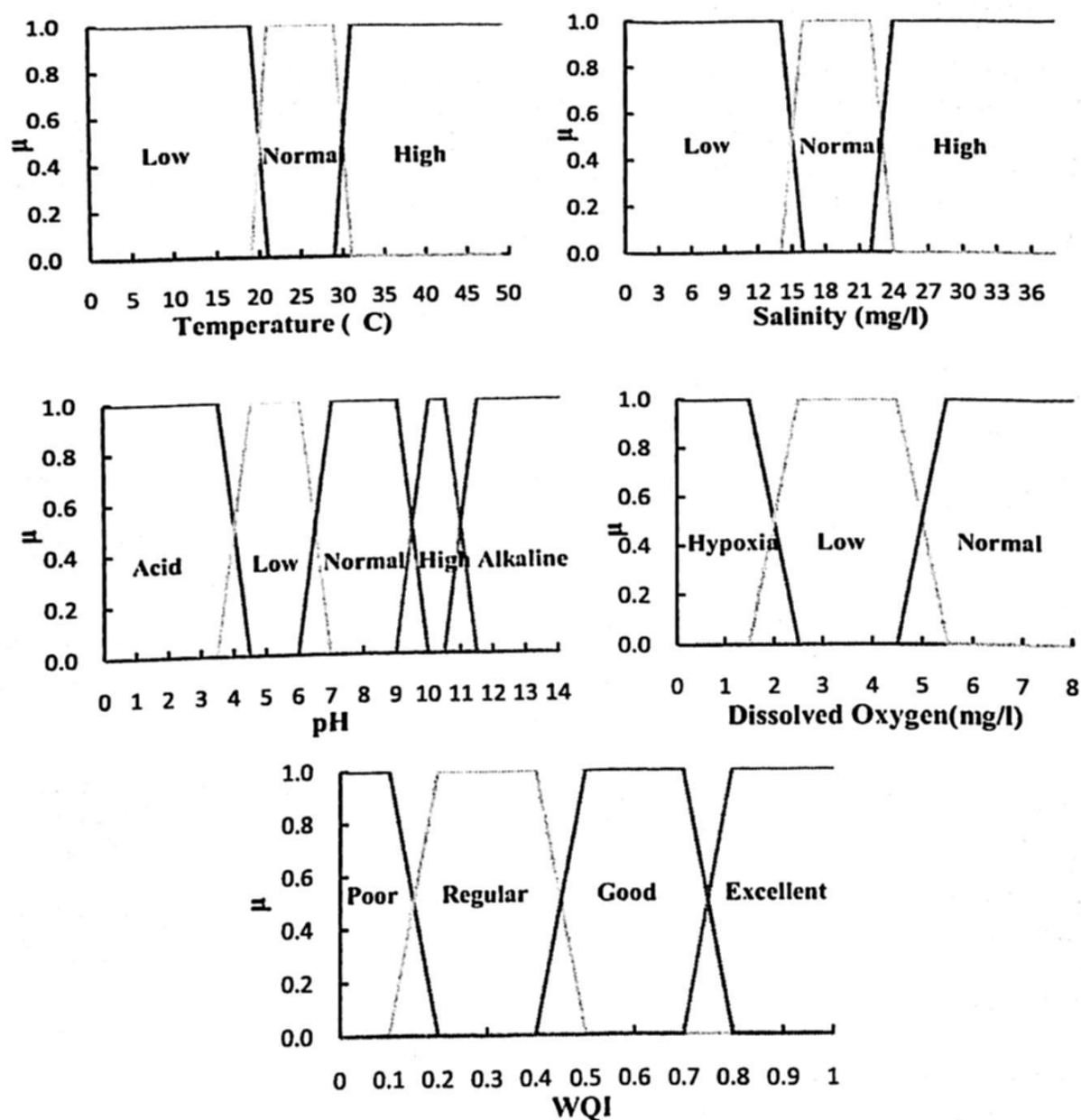


Fig. 4. Membership functions for temperature, salinity, dissolved oxygen, pH and WQI.

Water Quality status

The ecosystem is always changing, and the combination of the variable concentrations defines the status of the water quality. If a high impact variable reports harmful concentrations, therefore the status of the water quality will be deteriorated [1], [2], [3]. The water quality status has been classified in four levels, whose involve all the hypothetical situations in a shrimp pond:

1. *Excellent*: physical-chemical variables report concentrations in the optimal range.
2. *Good*: One variable reports concentrations out of the optimal range; however

- this situation do not represents danger in the shrimp.
3. *Regular*: some variables report concentrations out of the optimal range, and the combination between them represents certain stress level in the organism.
 4. *Poor*: all the variables concentrations are out of the optimal ranges, or a variable with a high impact level presents concentrations that could generate a potentially danger situation in the pond (p. ej. extremely low oxygen concentrations).

Reasoning process

There are some expressions that are frequently used by experts in water quality, that expressions will be helpful for the construction of the FIS. This kind of expressions implements the fuzzy language of the FIS and they are known as inference rules; they are represented as follows:

Rule 1: If Temp is *normal* and Salt is *normal* and pH is *normal* and DO is *normal* then WQI is *Excellent*

Rule 2: If Temp is *normal* and Salt is *High* and pH is *alkaline* and DO is *low* then WQI is *Poor*

The size of the set rule depends of the number of rules that are involved in the environment; a total of 139 rules have been used in this case. An inference rule process the membership functions values as:

$$\mu_R(\mu_{temp}, \mu_{pH}, \mu_{sal}, \mu_{DO}) = \min\{\mu_{temp}^i, \mu_{sal}^j, \mu_{DO}^k, \mu_{pH}^l\} \quad (6)$$

where i, j, k and l are the ranges of the evaluated variables respectively.

Aggregation

Output rules are matched with the WQI membership functions as follows:

$$\mu_{out} = \min\{\mu_R^l, \mu_{WQI}^l\} \quad (7)$$

where l defines the assessed status.

Finally all output rules are aggregated in order to create a final membership function, which is evaluated using a gravity center method [11]. It is used for transforming the output membership function in a $[0, 1]$ value, this value represents the Water Quality Index, where 0 means poor and 1 means excellent water quality. Gravity center is calculated using the following equation:

$$WQI = \frac{\int x \mu_{out}(x) dx}{\int \mu_{out}(x) dx} \quad (8)$$

Results using gravity center method never reach the maximum/minimum values, in order to rescale the output; the next expression allows having a $[0, 1]$ index.

$$WQI_n = \frac{WQI - \min(WQI)}{\max(WQI) - \min(WQI)} \quad (9)$$

where WQI_n is the normalized water quality index. The values for *excellent*, *good*, *regular* and *poor* are 1, 6.666, 3.333 and 0 respectively.

5 Software

A graphic user interface was developed in order to assess the artificial habitat in shrimp aquaculture and for an easier handling end-user. Software interface have a set of functions that allows the user assessing water quality, calculate averages, trends, save and restore information as main functions. LabVIEW was used as programming language, which is characterized for having a set of virtual instruments, with a graphic language and a customizable frontal panel as user interface [13].

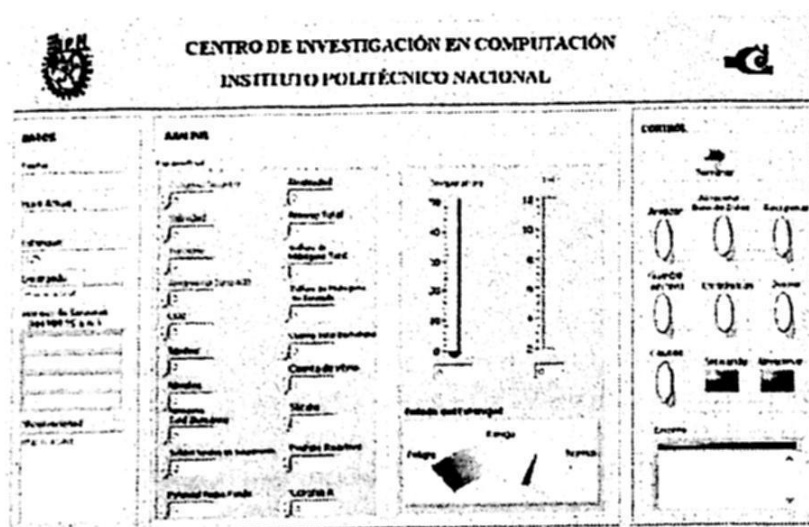


Fig. 5. User interface designed in LabVIEW environment.

6 Results

In order to prove the functionality of the fuzzy inference model, a data set was extracted from a database of a shrimp farm of Rancho Chapo located in Huatabampo, Sonora, which has a total 24 of measurements with a period of 15 minutes between measurements.

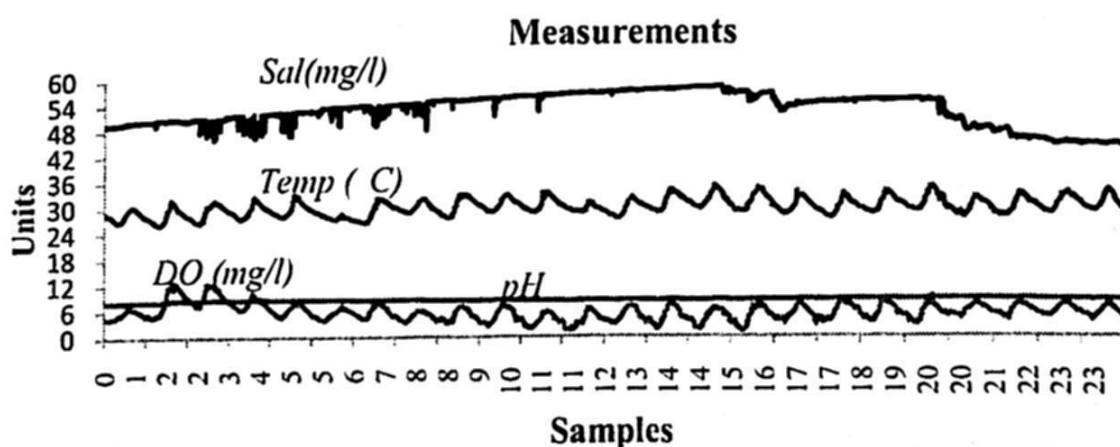


Fig. 6. Physical – chemical signals of the data set measurements registered in June of 2007.

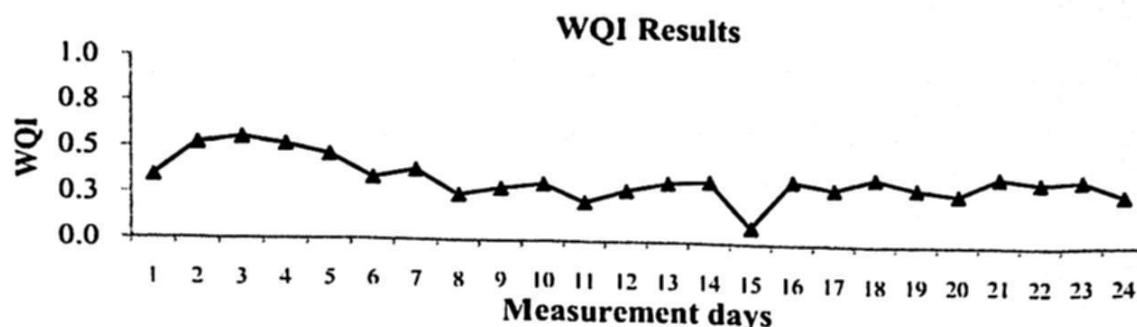


Fig. 7. Water quality assessment using the data set measurements of June of 2007.

Fig 6 and 7 shows a comparison between measurements and results; in Fig. 6 salinity reports values out of its range; Temperature reports values between 25 and 35 °C; DO reports oscillatory concentrations, whose are lightly above of normal range at day and bad range at night the firsts days. PH reports normal concentrations in the month.

In general salinity, temperature and dissolved oxygen report most of the time values out of range; it is clearly seen in Fig. 7, where most of the results are *regular* and *poor*. First days of June pH, temperature and dissolved oxygen concentrations are acceptable and the score is close to *good* water quality.

7 Conclusions

In this paper a distributed system for monitoring the water quality in shrimp farms has been developed. The distributed system is built in three phases; a) Sensor monitoring, b) data acquisition and c) signal processing. A set of four sensors have been conditioned form monitoring the most frequently measured physical-chemical variables. A Data acquisition card was designed receiving data sensors and for transmitting environmental information to a PC software. Assessment software was implemented using fuzzy inference system (FIS) theory for the processing of information. The distributed system results in an excellent tool for water management and treatment of the water quality in shrimp aquaculture.

Acknowledgementss

Authors wish to thank CONACYT, Centre of Biology Investigations of Sonora (CIB), Institute of Technology of Sonora (ITSON) and the National Polytechnic Institute (IPN) for supporting this work.

References

1. Angulo, C. A. y Angulo, C. U.: Estudio de Calidad del Agua y su Relación con el Crecimiento del Camarón Blanco (*Litopenaeus Vannamei*), en la Granja Camaronera Agua Verde, S.A. de C.V. en Rosario, Sin., Tesis, Universidad Autónoma de Sinaloa, 2003.
2. Hirano, Y.: Current practices of water quality management in shrimp farming and their limitations. Proceedings of the Special Session on Shrimp Farming. World Aquaculture Society, USA, 1992.
3. Martínez L.: Cultivo de Camarones Peneidos, principios y prácticas, AGT Editor, México (1994).
4. Li, Y., Li, J., Wang, Q.: The effects of dissolved Oxygen Concentration and Stocking Density on Grown and Non-Specific Immunity in Chinese Shrimp, *Fenneropenaeus Chinensis*. Aquaculture, Vol. 256. Elsevier (2006) 608-616.
5. Yew-Hu C.: Water quality requirements and management for marine shrimp culture. Proceedings of the Special Session on Shrimp Farming. World aquaculture Society, Baton Rouge, LA USA (1992) 144-156
6. Georg Fischer Signet LLC: Signet 2714-2717 Twist-Lock pH/ORP Electrodes datasheet www.gfsignet.com [Accessed February 2008].
7. Georg Fischer Signet LLC: Signet 2819-2823 Conductivity/Resistivity Electrodes datasheet www.gfsignet.com (2008).
8. Finesse, LLC:, Improved Process Measurement & Control: TruDO, Application note. www.finesse-inc.com, [Accessed February 2008]
9. PIC18F2550 Datasheet, Microchip Technology Inc, US, 2006.
10. Ocampo, W., Ferré, N., Domingo, J., Schuhmacher, M.: Assessing water quality in rivers with fuzzy inference systems: A case study. Environment International, Vol 2. Elsevier (2006) 733-742.
11. Rodríguez, A., Antonio, J.: Aplicaciones de lógica difusa en ingeniería gráfica, XVI Congreso Internacional de Ingeniería Gráfica (2004).
12. Fuzzy Logic Toolbox™ User's Guide, 13th Edition, The MathWorks, Inc. Natick, US, 2008.
13. Labview User Manual. National Instruments Corporation, US, 2003.

Intelligent Fault Diagnosis and Prognosis using state validations in a drinking water plant

Hector Hernandez¹, Jorge Camas¹, Nicolás Juárez¹, Madaín Pérez¹,
Rafael Mota¹ and Claudia Isaza²

¹Instituto Tecnológico de Tuxtla, ITTG
Carretera Panam. Km. 1080, 29050-Tuxtla Gutiérrez, Chiapas, México
(hhernandezd, jcamas, njuarez, mperez, rmota)@ittg.edu.mx
<http://www.ittg.edu.mx>

²Facultad de Ingeniería, Departamento de Ingeniería Electrónica
Calle 67 Número 53-108, Medellín, Colombia
cisaza@udea.edu.co
<http://www.udea.edu.co>

Abstract. This paper proposes a transitions validation method between system functional states in drinking water plant monitoring. The method is based in fuzzy entropy measure. The water plant is monitored by means of fuzzy classification method. Diagnosis of the system is the main objective of the work, nevertheless as a complement, prognosis is also proposed whenever maintenance takes an important place of the water plant. In this process, periodic maintenance is fundamental, and its schedule is commonly applied. Then, the objective of the study is coordinate intelligent fault detection with prognosis, in order to propose an adaptive and preventive maintenance schedule to achieve in integrated drinking water plant monitoring. This knowledge is then organized as a partition of the data set into classes representing the functional states of the process (normal or faulty operations). The proposed method validates the recognized functional state in presence of uncertainty using diagnosis and prognosis technique.

1 Introduction

Water industry is facing increased pressure to produce higher quality treated water at a lower cost. In drinking water industrial production processes, the correct operation and maintenance of the complex processes have a crucial role to ensure the supply of the quantity of adequate water to the population and the safeguarding of the environment. Monitoring principle of a dynamic process from a method of classification consists in determining at every moment, the current class which was associated beforehand a functional state of the system.

It's because that in the production phase (stage of recognition), it's a question of deciding sampling at every moment which is the operating condition. This decision is particularly delicate to take at the time of the transitions, i.e. when there is a change in the class to which the whole of measurements (individuals to be classified) is allotted.

The result of fuzzy classification techniques provides the individual adequacy degrees analyzed with each class. In the majority of the algorithms, the decision of classification is obtained by the research of the class to which the individual presents the maximum of membership, or adequacy. In the presence of uncertainties caused by the inaccuracy in measurements, or by the possible not very significant disturbances of the process operation, the transition from a state to another, it can to have a little real justification. In this case, we say that there is a "bad conditioning" to make the decision of change of state. This is why the introduction of a criterion of validation of the transitions was regarded as a significant contribution to the effective process monitoring.

This article approaches the validation of change of state to avoid false transitions or transitions in badly conditioned states. The validation is made in the stage of recognition. This approach is based on the information which each calculates from degrees of adequacy obtained by a fuzzy classification algorithm. For this approach, the essential criterion associated a decision was regarded as the evaluation of the information which was necessary to take it.

Generally, the quantity of information is associated the entropy of the data. In the case of fuzzy sets, the entropy nonprobabilistic formula suggested by Luca and Termini [2] is largely used. For transition validations, it is necessary to use the individual instantaneous information that produced the change of state. Consequently, the analysis is based on the unit (vector) degrees of adequacy of this individual to each class. The adequacy vector degrees can be regarded as a fuzzy unit. Transitions validation based on the quantity of information uses a traditional measurement of fuzzy entropy [2][3][4], the decision is considered well conditioned when the quantity of information is significant. However, a decision which was made with low values of the adequacy degrees gives a high value of information. For this, the paper propose to use a reliability index which was inspired by the measurement of fuzzy information (fuzzy entropy of Luca and Termini). This index allows measuring the instantaneous information which caused the change of state and allows holding in account the relationship between the small adequacy degrees and uncertainty on the decision. To validate the approach in experimental form, the method was applied to the process monitoring of drinking water plant of Tuxtla-Mexico. The practical significance of results will be in order to improve the water treatment process.

In section 2, this paper presents a brief description of water treatment process [5]. The transition validation method is presented in section 3, in section 4 the method is evaluated on a real case. Finally the conclusions are presented.

2 Water treatment process

2.1 Brief Drinking Water Plant description

The water is the most abundant compound on the surface of the world. Water treatment involves physical, chemical and biological processes that transform raw water into drinking water which satisfies a whole of standards of quality at a reasonable price for the consumer.

The "SMAPA" water treatment plant (Tuxtla city, Mexico)[10], which was used as an application site for this study, provides water to more than 800,000 inhabitants and has a nominal capacity to process 800 l/s of water per day. The figure 1 presents a schematic overview of the various operations necessary to treat the water.

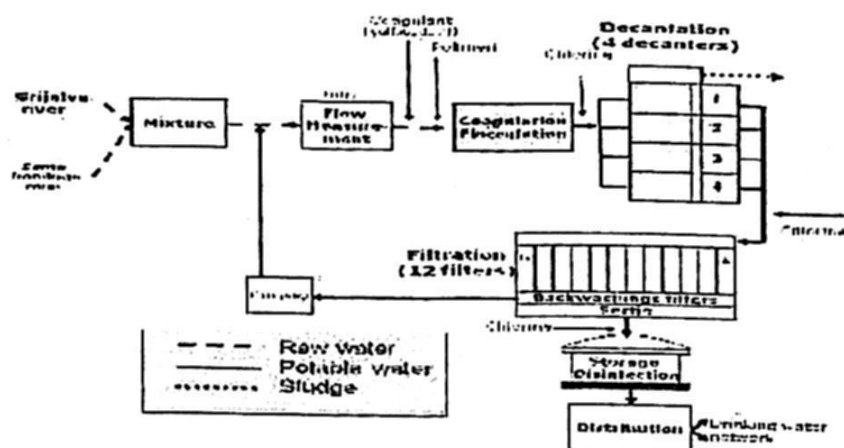


Fig. 1. SMAPA Potable Water Plant

The complete usual chain comprises the following units: clarification (coagulation-flocculation + decantation), disinfection, and filtration. The present work concerns essentially the clarification process which represents the main step in the treatment process. Raw water is collected at the rivers (Grijalva and Santo Domingo), and pumped to the treatment plants.

For decanters maintenance, the critical variable is the water turbidity at the input and the exit of the decanters (before filtering). The sludge accumulation in decanters is considered a non desired state of the plant. During the rainy season eventually there is a solids accumulation which blocks the sludge removal system. Consequently, a part of the solids is transferred to the filtering phase instead of being evacuated properly. Nowadays decanters maintenance is predetermined in February-March. The filters maintenance is programmed before the rain season so as to ensure the filter proper conditions at that period. As well, each filter has a backwashing task which is executed in function of the inlet and outlet filter pressure. A monitoring system of the process state would allow a flexible maintenance. Coagulation process is brought about by adding a highly ionic salt (aluminum sulphate) to the water. A bulky precipitate is formed which electrochemically attracts solids and colloidal particles. The solid precipitate is removed by allowing it to settle to the bottom of the tank and then periodically removing it as sludge. Water is stored in a tank and ready to be transported through the water supply network.

2.2 Monitoring Using Classification Methods General Theory

Process monitoring using classification method consists in determining at each sample time, the current class which was associated beforehand to a functional state of the process. There are two principal phases: the training and the recognition. In the

first step (training), the objective is to find the process behavior characteristics which will allow differentiating the process states (each one being associated to a class). The initial algorithm parameters are selected by the process expert who validates the obtained behavioral model. In a posterior step, the data recognition allows to identify one line the current process state. At each sampling time, a vector collects the accessible information (raw data or pre-treated data such as filtered, FFT) which is provided for monitoring, and the class recognition procedure yields the operator what is the current functional state of the process. In order to optimize the obtained partition we propose to include into the training phase a step to automatically validate and adjust the clusters. The proposed new approach automatically improves, in terms of compactness and class separation, a non optimal initial partition helping therefore the discrimination between classes i.e. between operation modes.

3 Transitions validation method

3.1 Fuzzy degree index (Decision index)

Entropy nonprobabilistic functions represent the fuzzy degree of a fuzzy discrete unit whole (μ) respect to the elements which make it up [2][6][7]. The analysis is made according to the adequacy degrees $\mu(x_i)$ of each element (x_i). According to the approach suggested by Luca and Termini, the entropy fuzzy functions type $H(\mu)$ must respond the following axioms [2][6][7][8]:

$$\begin{aligned} P1: & H(m) = 0 \Leftrightarrow m(x_i) \in \{0,1\} \\ P2: & \max H(m) \Leftrightarrow \forall i \quad m(x_i) = \frac{1}{2} \\ P3: & H(h) \leq H(m) \Leftrightarrow h \geq_s m \end{aligned} \quad (1)$$

The order relation \geq_s is a comparison operator called "sharpness". A fuzzy unit h is regarded as plus "acute" (sharp) than the fuzzy unit m if:

$$\begin{aligned} \forall x \in E \quad & \text{if } m(x) \leq 0.5 \text{ then } h(x) \leq m(x) \\ & \text{if } m(x) \geq 0.5 \text{ then } h(x) \geq m(x) \end{aligned} \quad (2)$$

Functions which obey these axioms can be expressed by the general formula:

$$H(m) = h \left(\sum_i^C w_i \cdot T(m(x_i)) \right) \quad (3)$$

C corresponds to the individual numbers in the dialogue universe (E) where is defined the fuzzy unit μ . According to [2][9]:

$$\begin{aligned} (i) & w_i \in \mathbb{R}^+ \\ (ii) & T(0) = T(1) = 0 \\ (iii) & T(m(x_i)): [0,1] \longrightarrow \mathbb{R}^+ \text{ it has a maximum value in} \end{aligned} \quad (4)$$

$m(x_i) = \frac{1}{2}$, and is monotonous for $m(x_i) < \frac{1}{2}$ and for $m(x_i) > \frac{1}{2}$

(iv) the function $h: \mathfrak{R}^+ \longrightarrow \mathfrak{R}^+$ is monotonous increasing

De Luca and Termini [2] proposed to use as function $T(\cdot)$ the Shannon probabilistic entropy [4] applied to the pair formed by the element and its complement to the fuzzy unit (equation 5):

$$T(m(x_i)) = m(x_i) \cdot \ln m(x_i) + (1 - m(x_i)) \cdot \ln(1 - m(x_i)) \quad (5)$$

Then the Luca and Termini entropy expression is [2]:

$$H_{DLT}(m) = K \cdot \sum_i^C S(m(x_i)) \quad (6)$$

where $K \in \mathfrak{R}^+$ is a normalization constant.

Many studies showed the validity of this expression like fuzzy information measures [1]. However, other families of functions with the same base form that the Luca and Termini can be used especially in the field of the decision-making and classification [11][12]. They are always based on the axioms presented in (1) and in the equation form of (3), where $w_i = K$ and the function T is given by:

$$T(m(x_i)) = f(m(x_i)) + f(1 - m(x_i)) \quad (7)$$

3.2 Validation index for states transition

Individual adequacy degree expresses his adequacy to a class. On the other hand, in the case of a decision-making between several groups or classes, the adequacy degrees express the adequacy of an individual to several classes. In the case of systems dynamic monitoring, the individual is represented by the whole of the variables which define the current state of the system and the classes are the possible states. The state with the highest adequacy degree is considered as the system that evolves at this time there. The reliability of the choice of the state at the moment present is directly proportional to the capacity of election among the adequacy degrees. On the contrary of fuzzy degree indices (e.g. fuzzy entropy), the problem in this case is not any more the analysis of adequacy of several individuals to a class, but the choice between several classes (states) to which an individual can belong. For this, this paper define a new fuzzy unit where the dialogue universe E is defined by the number of classes C and adequacy degrees correspond to the values of adequacy of each individual to each class. Plus one unit is ordered, plus informative is and then the entropy is lower. The unit that the paper consider most informative in the case of a choice is that which assigns the individual in a class with the maximum of adequacy while the adequacy degree of the other classes is null. On the other hand, the unit decision entropy becomes maximum if all adequacy degrees are equal. Consequently, the information provided by the unit would be then null, with respect to the reliability of the decision. In this context, as validation index this paper use the complement of the entropy of the decision suggested in [3] which is based on the Luca and Termini fuzzy entropy.

Considering that the choice corresponds to the maximum of adequacy, thus $\mu_M = \max [\mu(x_i)]$. Then, the fuzzy decision indexes are defined like the difference between this maximum value and each unit adequacy degree (case of state validations that corresponds to the individual adequacy degrees to each class):

$$d_i = m_M - m(x_i) \quad \forall i \neq M \quad (8)$$

The decision entropy for the fuzzy unit μ is the dual of total information (eq. (9)):

$$H_D(m) = 1 - I_D(m) \quad (9)$$

The useful information provided by the fuzzy unit whole for the decision-making $I_D(\mu)$ corresponds to:

$$I_D(m) = K \cdot \sum_i d_i \cdot e^{(d_i)} \quad \text{where} \quad K = \frac{1}{C \cdot m_M \cdot e^{m_M}} \quad (10)$$

This entropy is based on axioms which follow the philosophy of those proposed by Luca and Termini using the fuzzy entropy like fuzzy degree index. These axioms are:

$$\begin{aligned} \text{R1:} \quad & H(m) = 0 \Leftrightarrow \forall i \neq M ; \quad m(x_i) = 0 \\ \text{R2:} \quad & \max H(m) \Leftrightarrow \forall i, j ; \quad m(x_i) = m(x_j) \\ \text{R3:} \quad & H(\eta) \leq H(\mu) \Leftrightarrow \eta \geq_F \mu \end{aligned} \quad (11)$$

The relation \geq_F is proposed like a comparator of reliability to unit. This operator is defined in a way similar to the operator "sharpeness" ($=_S$) proposed by Luca and Termini. A decision based on a nonprobabilistic fuzzy unit h is considered more reliable than another based on the fuzzy unit m if the reliability index of h is larger than that of m [15].

$$h \geq_F m \Leftrightarrow FIA(h) \geq FIA(m) \quad (12)$$

Operator $FIA(\mu)$ is defined like:

$$FIA(m) = m_M + \text{card}[d(m)] \quad (13)$$

3.3 Transitions validations algorithm diagram

When a transition between the moment $t-1$ and the moment t is proposed by the recognition algorithm, this paper propose to analyze the information index (equation 10) of the new adequacy vector individual $x(t)$, if this one exceeds a certain uncertainty level the transition is validated. In the contrary case, the transition is put on standby, as long as the algorithm of recognition continues to propose the same class, it will be validated if at one posterior moment $t + r$ the quantity of information is considered reliable. The validation method diagram suggested is presented in Figure 2.

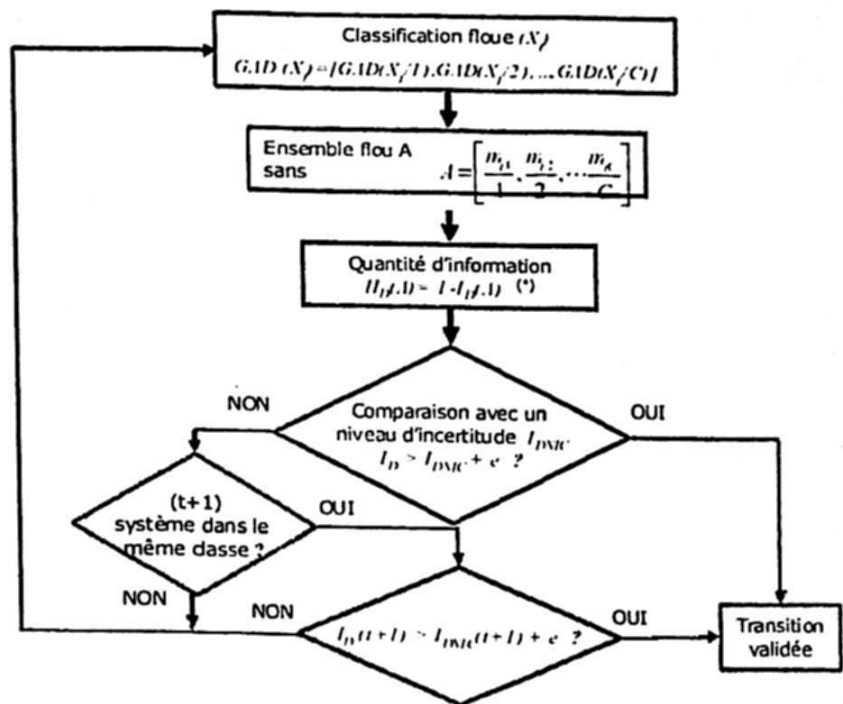


Fig. 2. Transitions validation algorithm diagram proposed

The information measure is the information index $ID(\mu)$ given by equation 10. It can arrive, either that a transition is not never validated or remains been unaware of system, or that it is validated with a delay R compared to the moment when the algorithm of recognition had detected it. The introduction of this delay makes it possible to eliminate the effects of noises or disturbances which provide an appearance of transition. In certain cases, there are oscillations between two classes which are only caused by the inaccuracies in measurements and these oscillations will then be eliminated in an automatic way, as well as false alarm which theirs is generally associated.

3.4 Uncertainty level

It is necessary to define a value minimum of information, which makes it possible to regard the decision as sufficiently reliable validating the change of state. This value can be established by analyzing the quantity of training information, in such manner according to the criterion of the system expert, the transitions considered as non valid will not be held in account during the stage from recognition. In this case there the uncertainty level is constant for all the time of supervision. However, it is more interesting automatically to obtain a value which corresponds to instantaneous minimal information to validate a transition at every moment from sampling. To obtain this value, on the LAMDA fuzzy classification method (Learning Method for Multivariate Data Analysis) [1][13], the information brought by attribution to one of the informative classes, (different from the class from Not Information, NIC), is equal to the uncertainty of this class NIC, since the union of the informative classes is exactly the complement of class NIC. The adequacy value of any individual to class

NIC is a constant provided by the algorithm which plays the minimum role of adequacy value, it is thus natural to use as uncertainty level (IDNIC) the formula of total information by including the adequacy degree of the class NIC, given by the equation 14.

$$I_{DNIC} = K \cdot (\sum_i (d_i) \cdot e^{d_i} + (d_{NIC}) \cdot e^{d_{NIC}}) \quad \forall i \neq M, \quad K = \frac{1}{C \cdot m_M \cdot e^{m_M}} \quad (14)$$

By analyzing the equation 14, if the decision change of class was made with an adequacy degree significantly higher than the minimal adequacy degree (μ_{NIC}), the relationship between the quantities of information corresponds to equation 15:

$$I_{DNIC} > I_D \Leftrightarrow m_M \geq m_{NIC} \quad (15)$$

To make the decision of validation this paper added a margin of decision e . This value makes the possibility to guarantee the information for the change of state ID and the minimal uncertainty level in such manner that the transition is valid if:

$$I_{DNIC} > I_D + e \quad (16)$$

If the decision margin is big, the transition will be validated and better conditioned.

With this approach, this paper validate only the transitions which have a sufficient information degree compared to total information, by including the class of minimal adequacy, which in the case of the method of classification LAMDA corresponds to the NIC.

4 Application to method to SMAPA potable water plant

The treatment plant concerned is the drinking water drinking water station "SMAPA" [10] of Tuxtla city in Mexico. For this station is significant to include an automatic monitoring system which makes it possible to have an alarm to avoid the states of bad operation. The objective is to establish a preventive maintenance according to the current state of the system and not on fixed dates like was made at the moment. For the monitoring system an approach of fuzzy classification was chosen, in such manner that according to the historical data a model station which allows, in the stage of recognition and in line can be established, to identify the states and to obtain an alarm to avoid the situations of bad operation.

LAMDA method was selected for data classification. The training time is very short and the parameters election method is easy to make for an expert in the procedure which does not have necessarily a strong knowledge on the classification methods [13]. To have a better performance and to retire the false alarms and conditioned bad classes, the transition validation method proposed in this document, was applied.

Training data base consists of 105 samples (from November 2000 to mid-February 2001). For each sample, the variable values which constitute a whole of 4 descriptors of the raw water quality are: turbidity input, a filter retrowashing number, plus the coagulant dose added. The fourth descriptor used is the difference between the value of turbidity input and the measured value after the decanters of the station. By using

the LAMDA classification method, the reference model is obtained made up of the functional states and alarms. After identifying the various classes, the system expert associates the classes to functional states and class was associates like alarm. The classification results (without the transition validations) which correspond at training stage are presented in Figure 3. The training is of the type not supervised, i.e. that the number of classes is not established a priori.

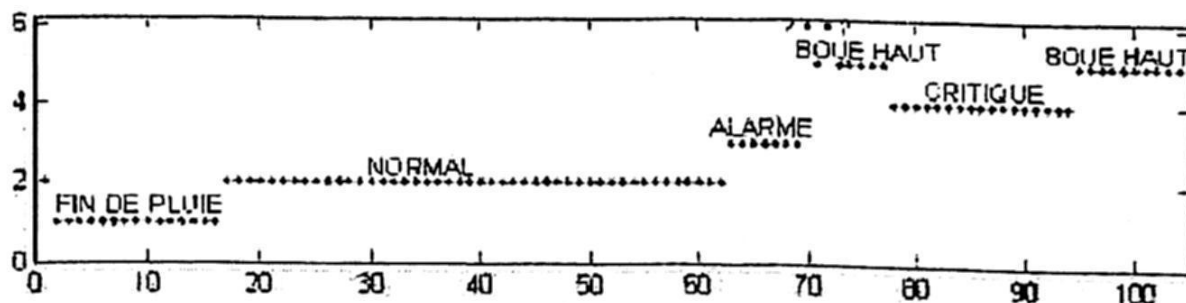


Fig. 3. LAMDA results: training phase without states validation (2000-2001)

Six classes were obtained. There are 2 classes which correspond to the normal operation, one of alarm which indicates the need for maintenance of the filters and two classes of faulty operation (high sludge, and critical stage). Class 6 is not a priori associated with a state with the system. This alarm which indicates a need for maintenance of the decanters intervenes very before the date established for maintenance. Indeed, for this whole of data the date of maintenance of the decanters (according to alarm) is proposed 87 days before the date plans. It's very significant to include this alarm to avoid high or critical sludge states (classes 4 and 5). During these states of bad operation, to give a quality of adequate water, it is necessary to increase the dose of coagulant and filter retrowashing number. If alarm is included, it is possible to carry out a preventive maintenance which will allow avoiding these undesirable situations.

To validate the classes and to retire the badly conditioned transitions, the transition validation method between states was applied. At every moment, the minimum level of information to validate a transition was calculated by holding of account the adequacy degree of each data with the class of noninformation NIC [13]. To utilize this method, it is appropriate to choose a value of requirement (e). With the whole of the training data, the value of requirement $e = 0.0018$ makes it possible to invalidate class 6. As it will be shown later, this value is representative of the system and not a priori of the training base. All the other transitions are validated perfectly. The result of the validation of transitions for this unit from data corresponds on Figure 4.

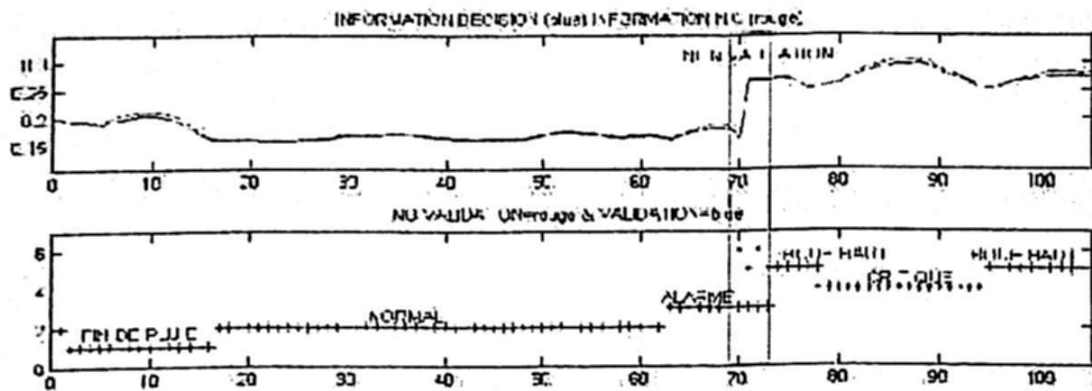


Fig. 4. States validation results: training data (2000-2001)

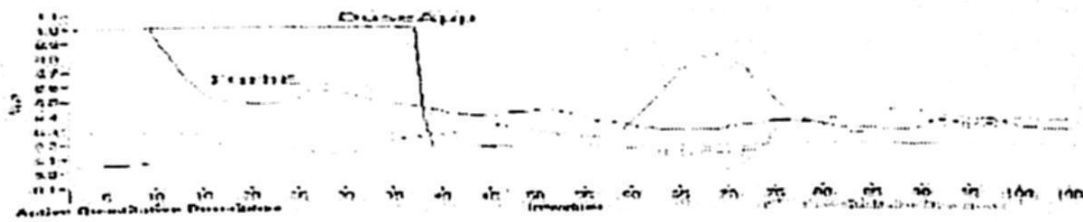


Fig. 5.a Analyzed variables, data test (2003-2004)

The variables which was analyzed (period 2003-2004) are presented in the figure 5.a. Classes identification results with the method LAMDA are watch in figure 5.b. and 5.c. shows the results applying this transition validation technique proposed. Class 6 was invalidated in all the cases, then it is regarded as state badly conditioned, alarm for the maintenance of the station was also identifies with new the data.

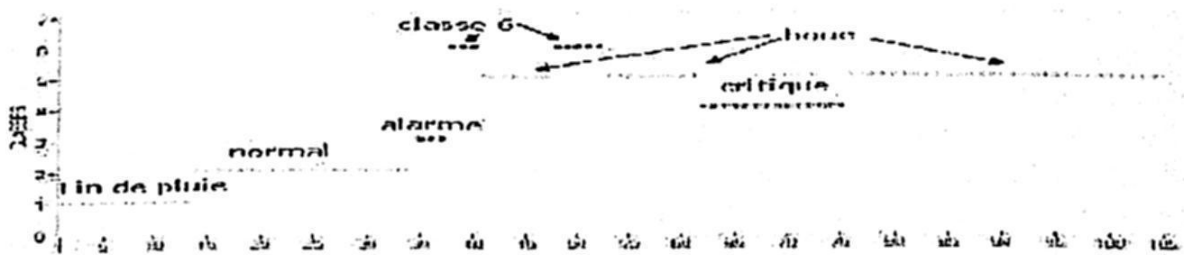


Fig. 5.b. Results from LAMDA: phase of test without validation of states (2003-2004)

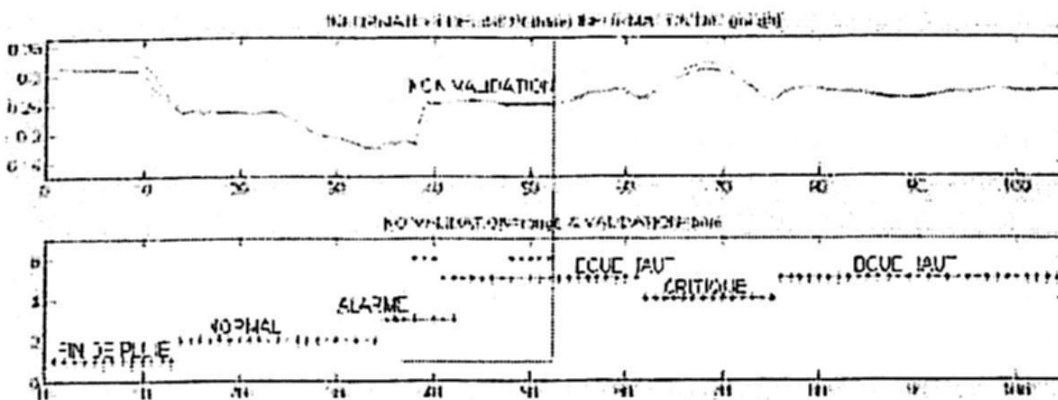


Fig. 5.c States validation results: data test 2003-2004

The method of validation suggested was also applied to a unit of data by replacing the dose of coagulant applied by the dose of coagulant calculated with neural networks [14]. In this case, there are false alarms which are removed with the transition validation method (Figure 6). The identification and validation of the system states for all the periods were made, without changing the value of the parameter e of validation method.

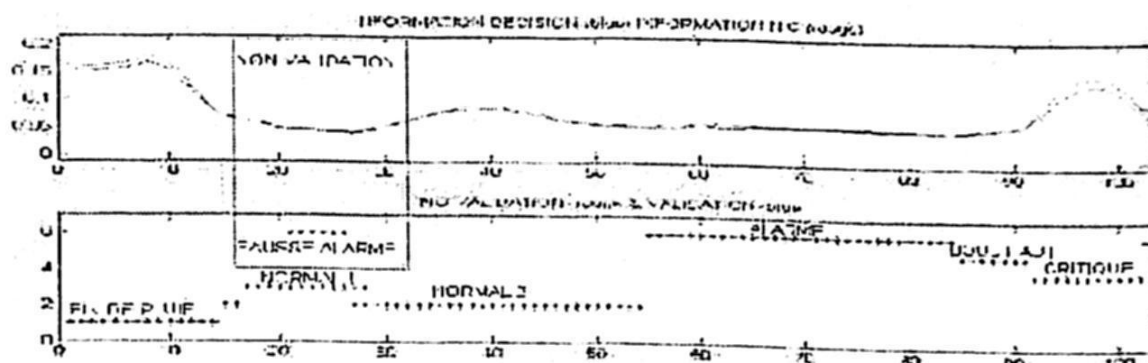


Figure 6. States validation results: data test 2001-2002 (Dose calculated)

In this case, there are three classes associated with normal operation (end of rains, normal 1 and normal 2). Class 6 corresponds to alarm but there is false alarm whereas one is in the normal state, this false alarm was eliminated by the method suggested.

5 Conclusions

The new methodology for transition validation based in fuzzy entropy measure is introduced. This strategy addresses the needs for diagnosis and prognosis in order to provide adequate preventive and predictive maintenance to potable water plants, which can lead to problems of dependability as well as significant economic losses. This approach provides a criterion for decision making when associating a class to an individual in presence of uncertainty or bad conditioned individuals. As a result, false alarms are eliminated.

Moreover, the effect of disturbances has been minimized when eventually they lead to non reliable transitions. In consequence, the system monitoring becomes more robust since apparent transitions due to inaccuracy measures are not validated. One of the advantages of the method is also that the transition may be validated since the method uses the output of the fuzzy classifier, which is not a big amount of data, and the computing time is as well reduced. On the other hand, some further studies are looking forward to introducing historical individuals memberships so that transitions could be validated based on finite time sliding window observations analysis.

References

1. T. KEMPOWSKY, A. SUBIAS, J. AGUILAR-MARTIN. Process situation assessment: From a fuzzy partition to a finite state machine, *Engineering Applications of Artificial Intelligence*, In Press, Corrected Proof, Available online 11 April 2006
2. D. LUCA and S. TERMINI. A Definition of Non Probabilistic Entropy in the Setting of Fuzzy Set Theory. *Information and Control*, 20:301-312, 1972
3. B. KOSKO. Fuzzy entropy and conditioning. *Info Sciences*, vol. 40, pp.165-174, 1986.
4. C. SHANNON. A mathematical theory of communication. *Bell SystTech*, pp379-423, 1948.
5. H. HERNANDEZ DE LEON. Supervision et diagnostic des procédés de production de l'eau potable. Thèse de Doctorat à l'INSA de Toulouse 2006. 04160 (2006).
6. E. TRILLAS, C. ALSINA. Sur les mesures du degré du flou. *Stoch*, vol.III pp. 81-84, 1979
7. E. TRILLAS, T. RIERA. Entropies in finite fuzzy sets. *Info Sciences* 15, pp.159-168, 1978
8. E. TRILLAS, C. SANCHIS. On entropies of fuzzy sets deduced from metrics. *Estadística Española* 82-83, pp. 17-25, 1979.
9. N. R. PAL and J. C. BEZDEK. Several new classes of measures of fuzziness. *Proc. IEEE Int. Conf. on Fuzzy Syst.*, 928-933, Mar. 1993.
10. SMAPA Sistema Municipal de Agua Potable y Alcantarillado de Tuxtla. *Manual de Procedimientos*, Tuxtla Gtz., Chiapas, México, 2007.
11. S. AL-SHARHAN et al. Fuzzy Entropy: a Brief Survey, *IEEE Int Fuzzy Syst Conf* 2001.
- A. DELUCA, S. TERMINI. Entropy of L-fuzzy sets. *Info and control*, 24, pp.55-73 1974.
- B. ISAZA, T. KEMPOWSKY, J. AGUILAR-MARTIN, M.V. LELANN, A GAUTHIE: « Etude comparative de la méthodologie LAMDA pour le diagnostic » ed. Cepadues , 2006.
12. H. HERNANDEZ DE LEON, M.V. LE LANN. Development of a neural sensor for the on-line prediction of coagulant dosage in a potable water treatment plant. *IBERAMIA* 2006.
13. E. DIEZ, J. AGUILAR-MARTIN; Proposition of Non-probabilistic Entropy as Reliability Index for Decision Making, *CCIA'2006*.

Web-Mapping Application to Retrieve Spatial Data by means of Spatial Ontologies

Miguel Torres, Marco Moreno,
Rolando Quintero and Giovanni Guzmán

Intelligent Processing of Geospatial Information Laboratory-Centre for
Computing Research-National Polytechnic Institute, Mexico City, Mexico
{mtorres, marcomoreno, quintero, jguzman!}@cic.ipn.mx
http://piig-lab.cic.ipn.mx/geolab, http://www.cic.ipn.mx

Abstract. Many types of information are geographically referenced and interactive maps provide a natural user interface to such data. However, the process to access and recover spatial data presents several problems related to heterogeneity and interoperability of the geo-information. We propose the **Tourism Onto-Guide-Web Application (TOGWA)**, which is a *web-mapping* system focused on retrieving geo-information by means of spatial ontologies and representing it on the Internet. Moreover, a Multi-Agent System is proposed to deal with the process related to obtain the tourist geo-information, which aids in the information-integration task for several nodes (geographic sites) that are involved in this application. The agent system provides the mechanism to communicate different distributed and heterogeneous Geographical Information Systems and retrieve the data by means of GML description. Also, this paper proposes an interoperability approach based on spatial ontologies matching. The matching is performed by the Multi-Agent System in every node considered in the application. The retrieval mechanism is based on encoding the information in a GML description to link the spatial data with the ontologies that have been proposed. TOGWA is a web-mapping system that is composed by two tiers: Client tier and Spatial Data Server tier, it offers an efficient and user-friendly interface to the clients.

1 Introduction

Maps are being used increasingly in local, networked and mobile information systems for communicating geographically referenced information. This has become possible because of the now relatively widespread availability of digital map data and developments in Geographical Information System (GIS) technology. The applications are widely ranging including local government planning, environmental monitoring, market analysis, navigation and public access to information [1].

Interaction with a digital map is typically based on a cycle of elicitation of user input via menu and dialog boxes, selection of map areas or features, and return of information, which may in turn induce modification to the map content. The maps themselves are often close replicas of traditional paper map cartography. The approach is to be found in many commercial GISs and is now being reflected in mapping applications on the Internet.

Developments in human computer interaction with regard to information retrieval and data visualization raise the question of whether the conventional approach can be improved. Certainly, there is a motivation to investigate new methods, since the current map interface, particularly on the Internet, often suffers from poor legibility of symbol and text, unnecessary user actions and inadequate adaptation to user interests [1].

Nowadays, the spatial databases are very useful and powerful tools to handle, display, and process the geographical information. These databases integrate GISs, which are composed to store and process spatial data. To solve some ambiguities in the spatial data processing and interpretation, the geo-information should have good quality from the input to the representation. The "adequate" representation of spatial data is crucial for improving the decision making process in different environments [2].

In this paper, we generate spatial ontologies based on the spatial semantics, which can be used to represent geographical objects by means of concepts ("not words"). Such spatial data conceptualization aims to compress the data and facilitate the knowledge discovery into spatial databases (SDB).

Up-to-date GISs do not extensively explore the spatial data semantics. To develop a spatial semantic theory is a great challenge in the new trends of Geocomputation field. Thus, the spatial analysis can use alternative methods to represent spatial data. This data representation jointly with the semantic rules - both based on data semantics - can be stored in a knowledge-base to generate new concepts that form the spatial ontologies. These concepts are defined by the properties and the behavior of geographical objects and explored by human experience. In general, we seek to correctly represent geographical objects for their subsequent processing [2, 3]. For instance, to retrieve spatial data from different SDBs and to represent them in the TOGWA.

In our proposal, the first step to generate the spatial ontologies is to obtain the *spatial semantics* of the geographical objects. By obtaining this definition, we can generate the spatial ontologies to rely on centralized ontology databases, which are stored in relational database systems. The emergence of Extensible Markup Language (XML) and Geographic Markup Language (GML) allows the ontology metadata to be embedded in the encoded web document, facilitating *semantic matching* by retrieval spatial concepts.

We propose a *Multi-Agent System* (MAS) to perform the communication between the different spatial databases. Although the encoding agents may still refer to centralized ontology databases during the encoding process, the spatial databases can also be encoded in GML because of its openness. Like many systems, we propose a Spatial User Interface Agent (SUIA) in TOGWA to make use of ontologies to validate user inputs and to capture the requests for retrieving spatial data by means of "concepts". In addition, the SUIA works in a web browser providing an easy-use web user interface for online geo-information retrieval. SUIA is characterized by the following features:

- Handle spatial data.
- Retrieve spatial data by means of concepts, considering the spatial subject domain.
- Perform spatial queries according to the generated spatial ontologies.
- Generate new geo-information making spatial analysis.

In this application, the spatial data are associated with different concepts, provided by the spatial ontologies. Moreover, an agent is considered to make several processes, which are divided into different tasks. The tasks that we are considering to it are the

following: representation of geographical phenomena, capability of communicating with the spatial data (different SDB), access to the spatial ontologies, and retrieval of the GML definition according to the user query.

The rest of the paper is organized as follows. In Section 2 we describe the Multi-Agent System proposed to perform several tasks in the application, and the spatial ontology to retrieve spatial data. In Section 3 we present the architecture of the TOGWA web-mapping application. The implementation of the prototype is shown in Section 4. Section 5 shows the preliminary results obtained by TOGWA. Our conclusions are sketched out in Section 6.

2 Multi-Agent System and Spatial Ontology

TOGWA is composed by two basis components to retrieve spatial data. These components are the following:

- A Multi-Agent System (MAS) to perform tasks related to communicate different spatial databases by means of GML definition and encode the spatial data for retrieving in the SUAI.
- Spatial ontologies to solve *ambiguities* that are presented in the spatial data through concepts ("not words").

2.1 The Multi-Agent System

According to [4], an agent is a system that tries to fulfil a set of goals in a complex, dynamic environment. It can sense the environment through its sensors and act upon it using its actuators. In this work, we propose a Multi-Agent System (MAS) that provides some services to facilitate the geo-information retrieval mechanism within a tourist system, which is called TOGWA.

There are two main functions of the agent in the TOGWA. One is to provide an intelligent service to communicate different spatial databases and to encode the spatial data for retrieving to the user. The other is to check the GML definition and to link the ontology for knowing whether the concepts accomplish to the search criteria.

Several types of agent have been proposed, they are organized in four layers depending on its functionality. The agents are shown in Fig.1.

- *Data Layer.* It is composed by the agents that provide data access services. These services can be: retrieval, storage, adjustment to the communication format, etc.
- *Management Layer.* The agents of this layer handle and coordinate other agents into MAS. Also, they provide the capabilities of communication to other agents.
- *Application Layer.* In this layer the agents perform tasks of the specific application, such as visualization and functions to the spatial data. Moreover, these agents manage the ontologies (*Trip Package* and *Map*) to provide the needed data to the interface for giving its own services.

- **Presentation Layer.** Here, the agents provide a user graphic interface to allow the users obtaining the TOGWA services.

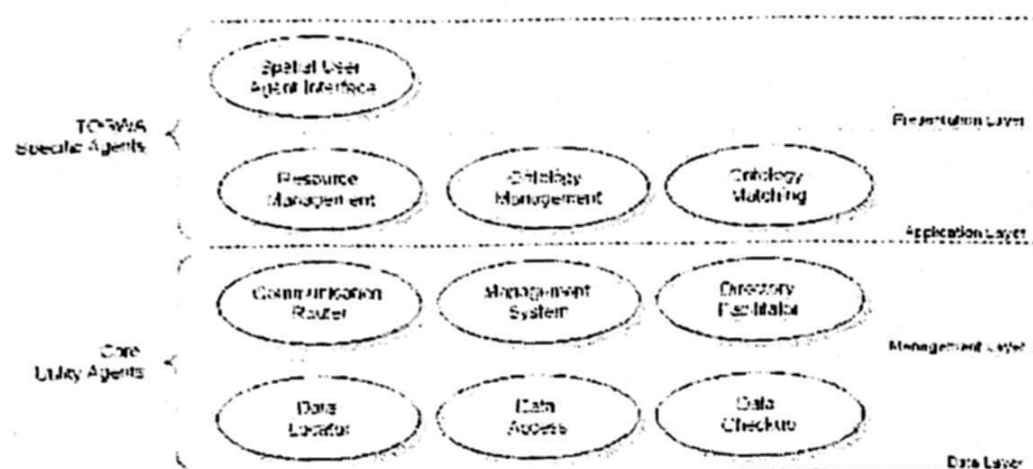


Fig. 1. Classification of the agents used in TOGWA

In Fig. 1 we grouped the layers in two clusters: Core Utility Agents and TOGWA Specific Agents. Core Utility is a set of agents that can be used not only in TOGWA but also in any other application. Its tasks are the following:

- **Data Locator.** It finds out the data that better fulfills the description given by its clients. The agent provides as a result the address of the agent, which can provide the access to the data.
- **Data Access.** It provides the mechanisms to access to the data and metadata of a particular source. The queries and results are given in XML.
- **Communication Router.** This agent provides the capabilities to MAS for communicating with other one, through any suitable way (Internet, others MAS, Virtual Private Networks, etc.).
- **System Management.** This agent handles the process within a MAS. It starts the compute of all other agents in the same MAS.
- **Directory Facilitator.** It maintains a list of all the known agents by MAS as well as the services that each agent provides to the layers.
- **Spatial Facilitator.** This agent retrieves the spatial data from the SDB. According to the client's request. The agent sends the geographical objects to make-up a map in the adequate format (GML description).

The TOGWA Specific Agents is a set of agents that work to accomplish specific TOGWA goals so that they can not be used in other applications. The agents that belong to this cluster are the following:

- **Resource Management.** This agent deals with all the resource assignment tasks. For instance, searching a hotel and flight and finding trip packages.
- **Ontology management.** This agent keeps the information about the Map Geo-Ontology and uses it to translate the user's request into structured queries, which will be computed into the Ontology Administration Query Module.

These queries allow other agents assigning resources to users and finding out geographical objects to provide maps to the clients.

- *Ontology matching.* This agent acts when there are confusions about the concepts in the client's ontology and the MAS ontology. Then the agent attempts to find the closest concept in MAS ontology, according to the concept given by the client.
- *Spatial User Interface Agent.* This agent translates data given by the MAS into a rendered map that the user can understand. The agent is the user interface of the MAS, but it does not belong to it.

The elements that compose the TOGWA-MAS are shown in Fig. 2. We can see that MAS consists of seven agents and an ontology repository.

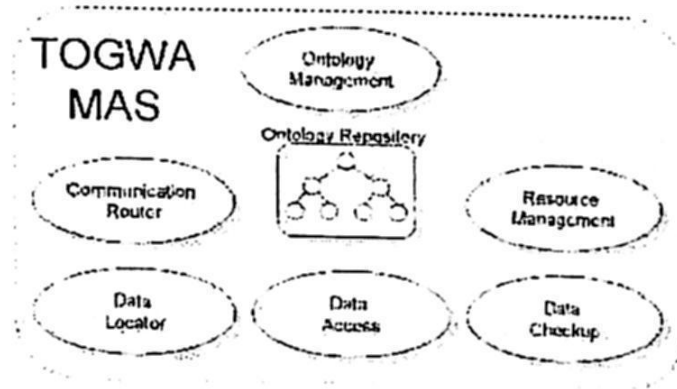


Fig. 2. Features that compose the TOGWA-MAS

In Fig. 3 the interaction model between two TOGWA-MAS is shown. Moreover, Fig. 3 shows the steps to accomplish the TOGWA process, which is outlined as followed.

1. The Client (Spatial User Interface Agent) makes a request to TOGWA (for example, a user in Spain desires to get a road map of the zone of Cancun in Mexico).
2. The MAS in Spain asks to its Directory Facilitator for the MAS that has such information.
3. The Directory Facilitator searches in its database the information requested, and responds to the MAS that the MAS of Mexico has the map.
4. The MAS in Spain asks the MAS in Mexico for the road map of the zone of Cancun.
5. The MAS in Mexico computes the request and determines if it has such information.
6. If it does, then the request passes to the Spatial Facilitator.
7. It makes a spatial query to the SDB for retrieving the geographical objects requested.
8. The Spatial Facilitator returns to the MAS in Mexico the geographical objects needed to make-up the requested map.
9. The MAS in Mexico translates this information using the ontology to a format that the MAS in Spain will understand.

10. Hence, it sends the information to the MAS in Spain.
11. Finally, the MAS in Spain sends the result to the client, and it displays the road map of Cancun to the user with a brief attributive description.

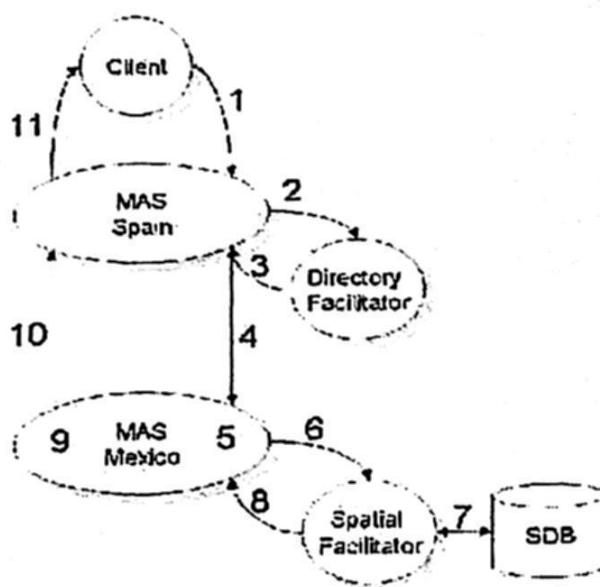


Fig. 3. Interaction model between two existing TOGWA-MAS

2.2 The Spatial Ontologies

Most widely accepted common conceptualization of the geographical data is based on the description of geographical objects and fields [5, 6]. These objects are not necessarily related to a specific geographic phenomenon, because human-built features are typically modeled as objects [7, 8]. The spatial semantics definition is described in [2] and aim to correctly represent spatial data in an alternative and universal way to generate spatial ontologies.

For this purpose, we consider a spatial ontology as an explicit, shared and structured specification of conceptualization, that is, a description of properties and relationships that can exist between the geographical objects to form concepts.

Moreover, ontologies can be considered as "languages", which use a specific vocabulary to describe entities, classes, properties and functions related to a certain view of the geographical world [9, 10].

In sense, our approach is designed to solve the ambiguities that can exist with single characteristics of the geographical objects, because the spatial ontology is defined by concepts (not by words) according to the geographical objects.

The spatial ontologies can be classified by levels according to their dependence on a specific task or point of view. These levels are generated for a specific spatial ontology (*top-ontology*) and it can be particularized to define a particular ontology (*down-ontology*). There are also different levels of information detail. *Low-level* ontologies correspond to very detailed information and high-level ontologies correspond to more general information.

In this situation, the generation of more detailed ontologies should be based on the *high-level* ontologies, so that each new ontology level can incorporate the new knowledge presented in the higher level. These new ontologies are more detailed, because they refine general descriptions of the level from which they have been generated [11].

The levels of ontologies can be used to guide processes for the extraction of more general detailed information. The use of multiple ontologies allows the extraction of information in different stages of classification.

The use of explicit spatial ontologies contributes to better correct spatial representation, because every geographical object description is based on an implicit ontology. By using that, it is possible to avoid explicit *conflicts* and *confusions* between the ontological concepts and the implementation [12].

On the other hand, spatial ontologies play an essential role in the conceptualization of spatial databases, allowing the establishments of correspondences and interrelations among the different domains of geographical objects and relations.

For instance, the ontology "Limit" can be represented in different concepts, in diverse spatial databases. "Limit" in some cases represents "coast boundary", separation between the "ground" and the "sea", "contour of value zero", "boundary" among two regions (states, countries, etc.), and so on (Fig. 4).

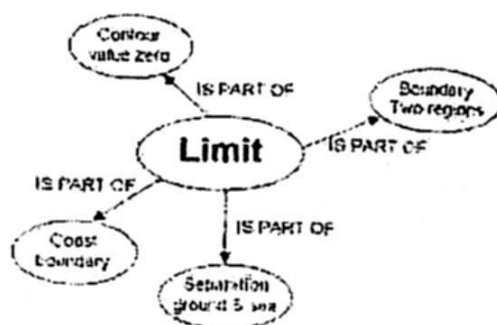


Fig. 4. Ontology "Limit" that is composed by different concepts related to "limit" for subsequent spatial representation

Using this approach, we can generate specific spatial ontologies after defining the top-ontology to particularize the conceptualization in other specific ontologies (down-ontology).

According to our approach, it is indispensable to count with a spatial subject domain. It is defined as a set of "names" that describe the primitives of spatial representation. Thus, we can start with *a priori* knowledge of the geographical objects that appear, e.g. in the map legend. For instance, "blue" lines are united under the concept (name) "river" and "black" lines are united under the concept "fracture", etc. In reverse, the different concepts are united under the same description of the spatial representation that is "line" [2]. The interaction between the subject domain and the taxonomy is used to locate concepts into the spatial subject domain that correspond to a case of study, and to compute these concepts to generate spatial ontologies [2, 4]. Fig. 5 shows the definition of spatial subject domain.

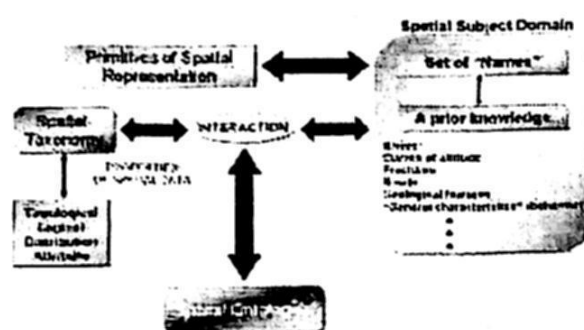


Fig. 5. Interactions of the Spatial Subject Domain

To define the spatial subject domain, it is necessary to elaborate a description of the spatial thematic to be analyzed, considering the main features that compose this theme, such as the data model and the resolution levels of the spatial data.

All characteristics that are considered in the description of the spatial subject domain should represent relationships between themselves too. Spatial subject domain has to recognize the different *semantic levels* of a *a priori* knowledge that is stored in this domain.

The use of ontologies in spatial databases enables knowledge sharing and information integration. The proposed approach provides dynamic and flexible information exchange and allows partial integration of spatial data when completeness is impossible.

This can help the next generation of spatial databases to solve *semantic ambiguities* in the available geo-information, because the context of the spatial data can change according to the case of study [11].

The query functionality to retrieve spatial data by means of spatial ontologies is the following: the user makes a request by means of TOGWA. The information is searched in any TOGWA-MAS, when the data are located; it is encoded into the GML definition. This definition is sent to the Ontology Administration Query Module to compute the request for obtaining the spatial and attributive data by means of concepts, which form the ontology. Inside TOGWA, *a priori* knowledge that is stored in the spatial subject domain interacts with the spatial taxonomy, considering in this case, the "arcs" as primitive of representation. Fig. 6 shows the mechanism to obtain the spatial ontology by means of the Ontology Administration Query Module.

In addition, Fig. 6 shows the query mechanism to describe the concept "Roads" into the *Map Geo-Ontology*. In this case, the ontology is composed by several "sub-concepts", which are ordered in a hierarchical way. Moreover, we see in Fig. 6 different levels of the concepts, starting with a top-level (Roads) and finishing with down-levels (One rail, two rails, etc.). When the ontologies present more levels of concepts, it is possible to particularize these in sub-concepts, while the level is less (down-level), the concept is more particular. The ontologies that have been proposed by TOGWA are described in Section 4.

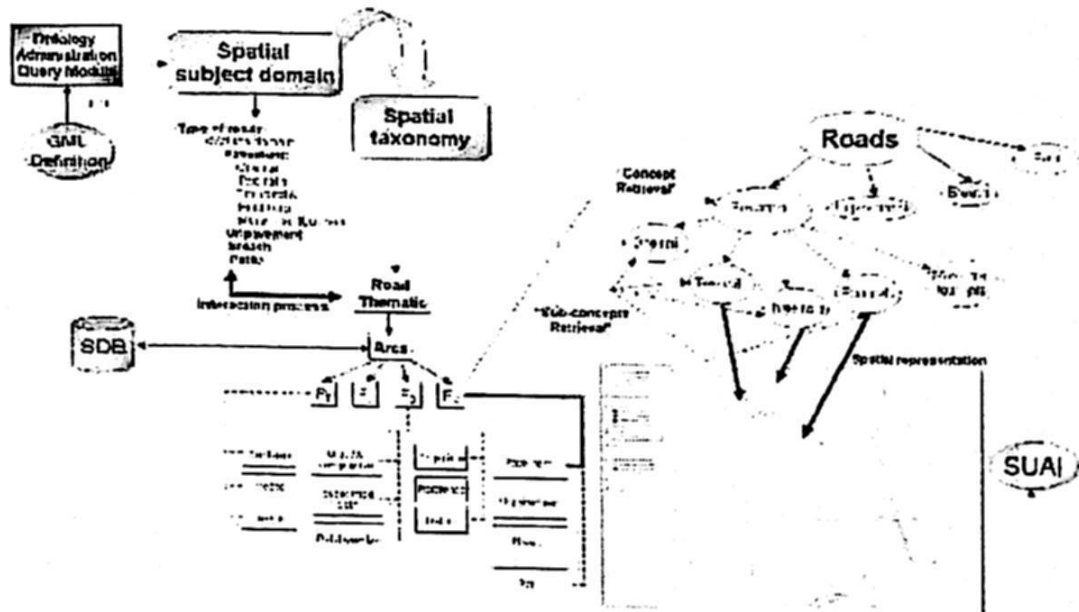


Fig. 6. Interactive process to retrieve geographical objects by means of spatial ontologies

3 Architecture of TOGWA

The Tourism Onto-Guide-Web Application (TOGWA) is a web-mapping system that is composed by two tiers: Client tier and Spatial Data Server tier. These tiers contain the following components: Spatial User Agent Interface (SUAI), Ontology Administration Query Module (OAQM), Spatial Data Server (SDS), Agent Administration Module (AAM) and Spatial Database (SDB) [13].

This application presents client-server architecture. TOGWA is considered a distributed system because it is able to retrieve spatial data from different GIS sites by means of GML definition. Fig. 7 depicts the general architecture of the Web-Mapping system.

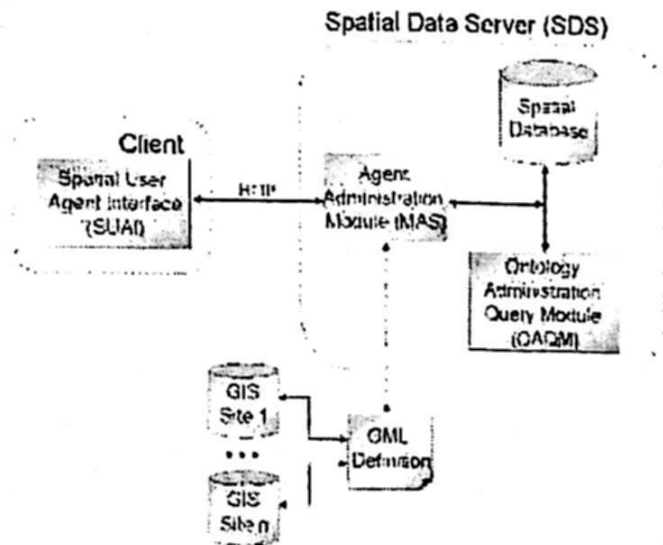


Fig. 7. Architecture of the Tourism Onto-Guide-Web Application

The general process to retrieve spatial and attributive data is the following:

Spatial User Agent Interface (SUAI) receives requests from a user. It assists the client to search, query and manipulate the map in an efficient and user-friendly way. SUAI attempts to understand the subject domain (geographical context of the user), and it sends a message to the Spatial Data Server (SDS) to ask more geo-information or to modify the map to change the content and resolution detail. SUAI should keep a concise profile for each user to record his search of interest. The Agent Administration Module (AAM) receives requests from the SUAI and it broadcasts the requests of the users to the Ontology Administration Query Module (OAQM) to search the concept into the ontologies and to retrieve the geo-information from the Spatial Database (SDB). If the geo-information associated to the concept could not be found in the SDB, the OAQM will send a notification to the AAM to perform a query in different GIS sites linked to TOGWA. This process is made up by means of the GML definition, when the geo-information is found; it is encoded in the GML description and transferred to the AAM to retrieve the spatial data according to the spatial ontology. Finally, the spatial data is sent to the SUAI.

4 Implementation of the Prototype

TOGWA prototype has been implemented in Java to keep the distribution and multi-platform execution [14]. TOGWA consists of seven nodes to retrieve spatial and attributive data. The nodes that are considered to this application are the following: Mexico, Spain, Costa Rican, Italy, England, Cuba and Chile. The SUAI is implemented as a Java Applet and runs on the client side to interact with a web user. The AAM has been implemented as a Java servlet using Tomcat 5.0.12. The visualization on the client side is based on *Shapefiles*, which is proposed by Esri, Inc [15].

There are several components in the prototype, a web page, an invisible applet, a servlet and an ontology parser. The data workflow is depicted in Fig. 8.

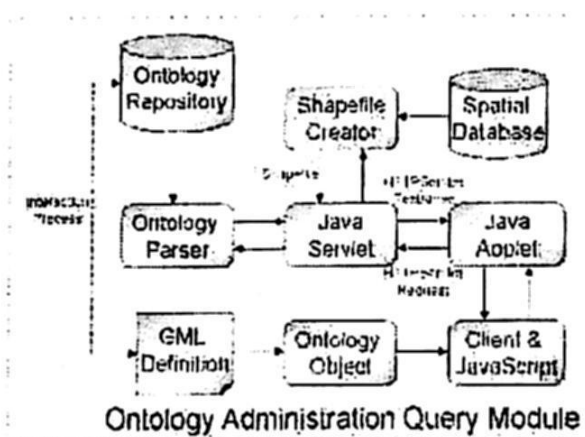


Fig. 8. Data workflow to obtain the spatial data according to the spatial ontology

When a user accesses the web page, the JavaScript embedded in the web page will call a Java applet to send an http request to the Java servlet, which will invoke the ontology parser to create an ontology object from the ontology repository. If the information does not find in the Ontology Repository, the OAQM sends a GML definition to locate the data in any node. When the information is found, it is received by the OAQM for being computed. Later, the OAQM sends the object as a serialized Java object to the applet.

The ontology object contains the entire ontology. The applet uses the ontology object to verify if the user has performed a valid search. If valid, the applet will submit the search to the servlet, which in turn invokes the shapefile generator to obtain a shapefile for the client to refresh the web page and to retrieve the spatial data.

In this context, a spatial ontology is a part of knowledge, concerning a particular spatial subject domain; it describes a spatial taxonomy of concepts for that subject domain, which define the *spatial semantic interpretation* of the knowledge. Spatial ontologies in TOGWA define the *spatial semantic relationships* of the geographical objects. The ontology repository is organized in a tree structure.

We propose two ontologies to obtain the spatial data by means of concepts, in this case the *Map Geo-Ontology* and *Trip Package Ontology*.

These ontologies provide the concepts related to the information retrieval to the user. The retrieval process is performed by constraints, which are defined by the client.

Map Geo-Ontology is focused on retrieving particular maps of the user interest. This spatial ontology can generate four types of maps: Roads, Weather, Urban and Sightseeing. The spatial ontology is generated by the interaction process of the spatial taxonomy and the spatial subject domain.

Trip Package Ontology is proposed to acquire attributive data related to the interest places to visit for the users. A user can obtain relevant information according to his interest and the matching concepts in the definition. The data can be found in any node considered into the application.

The interaction and communication process has been described in previous section (Multi-Agent System). The retrieval of concepts among ontologies is performed considering the *relationships* of them by means of MAS. Fig. 9 and Fig. 10 show the ontologies that we propose to retrieve spatial data by means of concepts.

On the other hand, the GML definition is used to obtain the spatial data from different distributed GIS according to the request of the user. MAS sends this definition to find the specification related to the request. If the data has been found, the GML definition encodes the information, which is sent to the Ontology Repository for matching this information encoded into the GML definition with the ontology structures. Inside the Ontology Administration Query Module the information is parsed for relating it with the concepts, which integrate the ontologies. We use the *relationships* between concepts that belong to the ontology to communicate the *Map* and *Trip Package Ontologies*. The use of relationships of concepts provides detailed information (spatial and attributive), because we can obtain concepts in certain directions as breadth and depth first search.

If the information is valid, it is necessary to generate the *shapefile* by means of the Shapefile Creator for sending it to the Spatial User Agent Interface. A brief description of the GML definition is shown in Table 1.

Table 1. Brief GML description related to TOGWA

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xml:lang="en"
  xmlns:camb="http://www.w3.org/2001/XMLSchema#dateTime"
  xmlns:gml="http://www.opengis.net/gml"
  >
  <!--The camb and gml namespaces replaced for validation purposes Map Ontology Data-->
  <camb:Map>
    <gml:boundedBy>
      <gml:Box gml:srsName="ROAD:4326">
        <gml:coordinates>
          0.0,0.0 100.0,100.0
        </gml:coordinates>
      </gml:Box>
    </gml:boundedBy>
    <camb:modelDate>
      Dic 2003.
    </camb:modelDate>
    <camb:modelMember>
      <camb:Roads>
        <gml:name>
          I45
        </gml:name>
        <gml:description>
          Federal Highway from San Pablito to Cancun.
        </gml:description>
        <gml:centerLineOf>
          <gml:LineString gml:srsName="ROAD:4326">
            <gml:coordinates>
              0.0,50.0 100.0,50.0
            </gml:coordinates>
          </gml:LineString>
        </gml:centerLineOf>
        <camb:ModelMember>
          <camb:Highway>
            <gml:Name>
              Interstate 35
            </gml:Name>
            <gml:description>
              Main Highway to connect Cancun and Chetumal.
            </gml:description>
          <gml:LineString gml:srsName="ROAD:4326">
            <gml:coordinates>
              344,552.4,566,763.67,763,234.12, 3,456,655.65, 890,765.31
            </gml:coordinates>
          </gml:LineString>
        </gml:centerLineOf>
      </camb:Roads>
    </camb:modelMember>
  </camb:Map>
</rdf:RDF>
```

```

</gml:LineString>
  </camb:Highway>
    </camb:ModelMember>
  </camb:Roads>
</camb:modelMember>
</camb:CityModel> ... </rdf:RDF>

```

5 Preliminary results by TOGWA

By using TOGWA, we have developed roads, city, weather and sightseeing maps. These maps are generated by means of concepts that belong to the *ontologies* (*Trip Package* and *Map*). The data have been retrieved by the GML definition according to the user request. *SUAL* contains an efficient and user-friendly interface, which is composed by some spatial tools. Some results are shown in this section.

Fig. 11 depicts the *map of roads* for Toluca City, Mexico. This map consists of different thematics as Populations, Roads, Urban Areas and Internal Administrative Divisions. The roads presented in this map are classified by its properties: four rails, two rails, tracks and urbanized routes. The goal of this map is to guide the users for knowing their interest places in low level of detail (1:200,000).

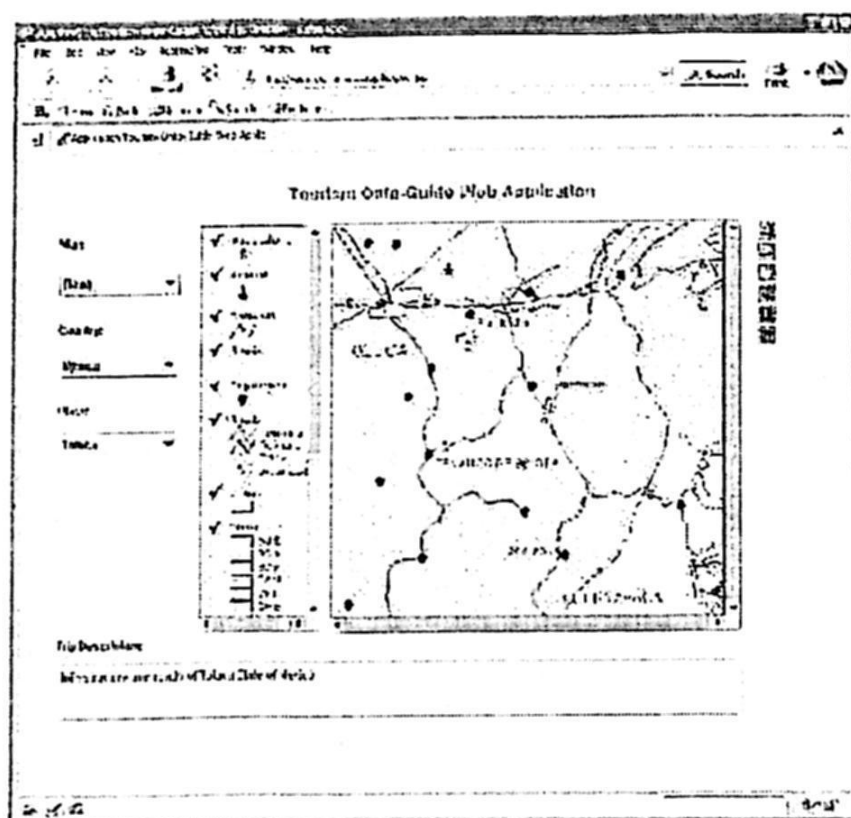


Fig. 11. Map of roads

City map is composed by streets, avenues and present different interest sites. Fig. 12 shows the city map of Lindavista area in Mexico City, which scale is 1:5,000. In this

map we show the location of different sites as Restaurants, Bus Stations and Hotels in this area. The users can retrieve a *city map* according to their necessities. Also, *Trip Description Box* provides useful information related to the user request.

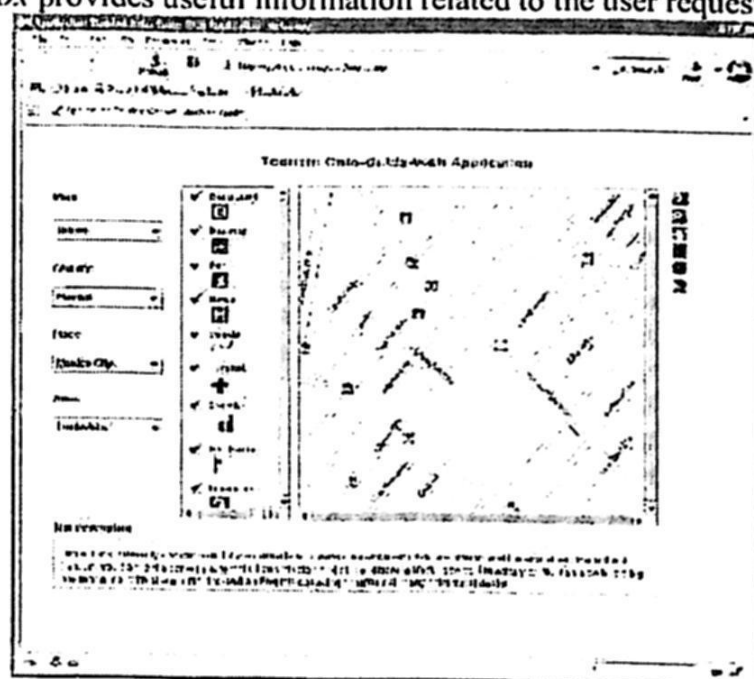


Fig. 12. City Map

Fig. 13 shows the *Sightseeing map* of San Pablito in Quintana Roo, Mexico. This map describes general aspects of San Pablito, showing the Information Sites Location, Gas Stations, Camping Zones, Restaurants and Archeological Sites. Moreover, it provides the general structure of the population. Additionally, this map presents the roads that connect with San Pablito (in red color). The map scale depends on the size of the interest area.

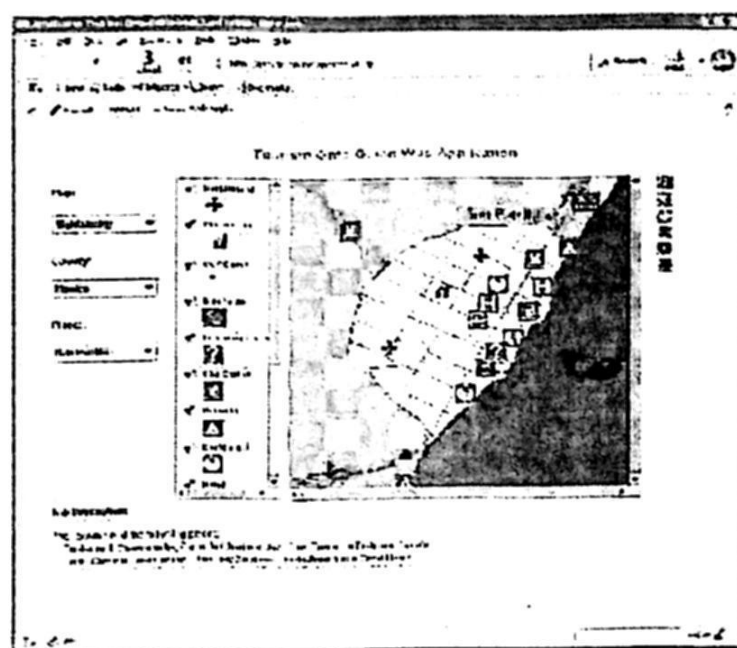


Fig. 13. Sightseeing Map

Weather map consists of vegetation areas, temperature and precipitation contours. This map guides to the users to know the characteristics of the weather in a particular place, when the users want to travel according to their criteria of retrieval request. In addition, Fig. 14 depicts attributive information related to the map into the *Trip Description Box*.

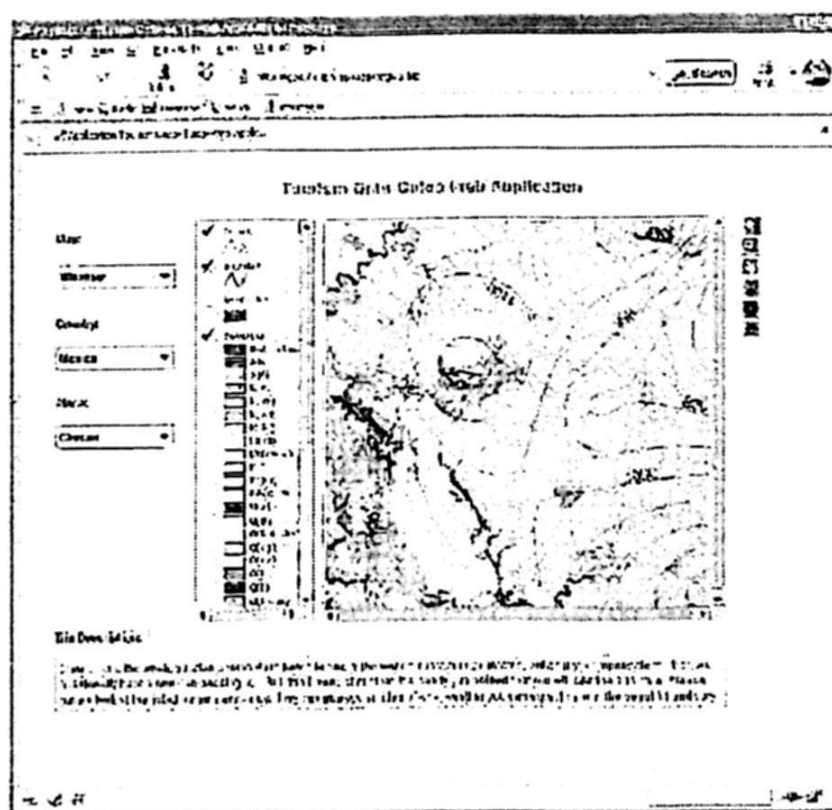


Fig. 14. Weather Map

6 Conclusions

In the present work, the Tourism Onto-Guide-Web Application (TOGWA) has been proposed. TOGWA is a *web-mapping* system focused on retrieving geo-information by means of spatial ontologies and representing it on the Internet. We use the spatial semantics to generate the geo-ontology for representing geographical objects by means of concepts.

TOGWA contains a Multi-Agent System, which performs the following tasks:

- To communicate different spatial databases by means of GML definition.
- To encode the spatial data for retrieving in the SUAI.
- To solve *ambiguities* that can be presented in the spatial data by means of concepts ("not words").

The spatial subject domain definition is oriented towards an interaction with spatial taxonomy to conceptualize the spatial databases. In essence, the spatial subject domain is defined as a set of "names" that describe the primitives of spatial representation. Thus, we can start with *a priori* knowledge of the geographical objects to examine the spatial data, which interact with the spatial taxonomy to generate spatial ontologies.

We attempt to show an alternative approach to represent spatial data on the Internet considering the *relationships* that compose the ontologies to retrieve spatial data according to several search criteria.

In addition, the spatial ontologies catch the semantics of the spatial data to provide relevant information related to the concepts. These ontologies can be used to establish agreements on diverse views of the world and consequently to carry out the "meaning" of the geo-information. In many situations, this geo-information is embedded in the spatial representation of geographical phenomena in the human-mind.

The use of ontologies in spatial databases enables knowledge sharing and information integration. The proposed approach provides dynamic and flexible information exchange and allows partial integration of spatial data when completeness is impossible in the web.

The communication between ontologies is performed by MAS, which seeks the relationships of the concepts to match nodes in the ontologies. This process is iterative and the new generated concepts can be considered in the spatial subject domain.

This approach can aid to solve *semantic ambiguities* between the available geo-information, because the context of the spatial data can change, according to the case of study and the representation state by means of concepts of the spatial data.

Acknowledgments

The authors of this paper wish to thank the SIP and CONACYT Projects: 20101069, 20101282, 20101088, 20100371, 106692; IPN and CONACYT for their support.

References

1. Li, M., Zhou, S. and Jones, C.B.: Multi-agent Systems for Web-Based Map Information Retrieval. In Egenhofer, M.J. and Mark, D.M. (Eds.), *GIScience 2002, Lecture Notes in Computer Science* Vol. 2478 (2002) 161-180
2. Torres, M. and Levachkine, S.: Semantics Definition to Represent Spatial Data, In: Levachkine S., Bodanský E. and Ruas A., (eds.), *e-Proceedings of International Workshop on Semantic Processing of Spatial Data (GEOPRO 2002)*, Mexico City, Mexico (2002)
3. Egenhofer, M. and Frank, A.U.: LOBSTER: Combining AI and Database Techniques for GIS. *International Journal of Photogrammetric Engineering and Remote Sensing*, Vol. 56, No.6 (1997) 919-926
4. Maes, P.: Modeling Adaptive Autonomous Agents. *Artificial Life*, No. 1 (1994) 135-162
5. Fonseca, F. and Egenhofer, M.: Ontology-Driven Geographic Information Systems, *Proceedings of 7th ACM Symposium on Advances in Geographic Information Systems*, Kansas City, United States (1999) 14-19
6. Fonseca, F., Egenhofer, M. and Agouris, P.: Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS*, Vol.6, No. 3 (2002) 25-40
7. Egenhofer, M. and Frank, A.: Naive Geography, in Frank A. and Kuhn W., (Eds.) *Spatial Information Theory, A Theoretical Basis for GIS, Proceedings of the International Conference COSIT '95, Lecture Notes in Computer Science*, Vol. 988, Springer-Verlag, Berlin (1995) 1-15
8. Gooldchild, M.F., Egenhofer, M., Fegeas, R. and Kottman, C. *Interoperating Geographic Information Systems*, Editorial: Kluwer Academic Publishers (1999)
9. Guarino, N.: Formal Ontology and Information Systems, in Guarino (Ed.), *Formal Ontology in Information Systems*, Editorial: IOS Press (1998) 3-15
10. Guarino, N.: Formal Ontology. Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies*, 43, Vol. 5, No. 6 (1999) 625-640
11. Guzmán, A., Domínguez, C. and Olivares, J. Reacting to unexpected events and communication inspite of mixed ontologies, In: C. Coello, A. Albornoz, L. Sucar, O. Cair and G. Kemper (eds.), *Advances in Artificial Intelligence, Proceedings of Mexican International Conference on Artificial Intelligence (MICA 2002)*, Vol. 2313, *Lecture Notes in Computer Science*, Springer-Verlag, Merida, Yucatan, Mexico (2002) 377-386.
12. Levachkine, S. and Guzmán, A.: Relatedness of the elements of hierarchies partitioned by percentages, *Lecture Notes in Computer Science*, Vol. 2972, Springer-Verlag, Berlin (2004) 135-155
13. Torres, M., Moreno, M., Menchaca, R. and Levachkine, S.: Making Spatial Analysis with a Distributed Geographical Information System, *Proceedings of 21st IASTED International Conference on Databases and Applications (DBA' 2003)*, Innsbruck, Austria (2003) 1245-1250
14. Zhou, S. and Jones, C.: Design and Implementation of Multi-Scale Databases, *Proceedings of 7th International Symposium on Spatial and Temporal Databases (SSTD01)*, *Lecture Notes in Computer Science*, Vol. 2121, Springer-Verlag, Berlin (1995) 365-384
15. <http://www.esri.com>, Environmental Systems Research Institute

Applying Diverse Data Mining Methods in the Electric Power Industry

Manuel Mejía-Lavalle, Guillermo Rodríguez O.,
Gustavo Arroyo F. and Eduardo F. Morales¹

Instituto de Investigaciones Eléctricas, Reforma 113, 62490 Cuernavaca, Morelos, México
¹ INAOE, L.E. Erro 1, 72840 StMa. Tonantzintla, Puebla, México
{mlavalle, gro, garroyo}@iie.org.mx
emorales@inaoep.mx

Abstract. We present our experiences in four Mexican power electric industry domains where we applied diverse data mining techniques. The first domain is about electric generator diagnosis. The second one is related to flashover forecasting in high-voltage insulators. The third case is about obtaining expert knowledge, applying data mining techniques to hydroelectric and thermoelectric utilities databases. The last case approaches a pattern recognition problem to detect potential electric illicit users. We outline successful and bad practices, and comment about possible solutions or future work that we think it have to be done to maximizing the usefulness of the data mining approach.

1 Introduction

Data mining has been employed with success in various fields and in many real world problems [1]. Data mining is applied to huge volumes of historical data mainly with the expectation of finding knowledge, or in other words, when it is sought to determine trends or behavior patterns that permit improve the current procedures of marketing, production, operation, maintenance, or others. In summary data mining, or more widely expressing, knowledge discovery, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [2]. Some of the traditionally used computer techniques to accomplish data mining are: neural networks, induction of decision trees, decision rules and case-based methods.

In this paper we present our experiences in several Mexican power electric industry domains where we applied diverse data mining techniques. The first domain is about electric generator diagnosis using expert systems plus a novel neural network paradigm. The second one is related to flashover forecasting in high-voltage insulators, where we proposed several tools to approach this problem. The third case is about obtaining expert knowledge, applying and comparing well known data mining techniques to hydroelectric and thermoelectric utilities databases. The last case approaches a pattern recognition problem to detect potential electric illicit users, where we proposed and realized a pre-processing feature selection method. We outline successful and bad

practices, and comment about possible solutions or future work that we think it have to be done to maximizing the usefulness of the data mining approach.

2 Electric Generator Diagnosis

In this Section we present an Electric Generator Failure Diagnosis (EGFD) system. The EGFD system combines two artificial intelligence approaches: expert systems and artificial neural networks. With our expert system shell we capture the human expertise. With our neural net paradigm we obtain knowledge from data. For instance, human experts on electric generation failures know that:

- a) Internal partial discharges can occur within the ground wall insulation at delaminations or areas where the bonding material is missing or incompletely cured.
- b) Such discharge activity is particularly common in older insulation systems such as mica folium and asphalt-mica.
- c) The main characteristic of this mechanism is that the positive and negative partial discharge activity is about equal.

Then a production rule to detect this condition becomes as follows:

RULE 3

IF: - Insulation system is made of <mica folium> or <asphalt-mica>.

and - [Positive PD at 50 mV] * 1.1 >= [Negative PD at 50 mV].

and - [Positive PD at 50 mV] * 0.9 <= [Negative PD at 50 mV].

and - [Positive PD at 200 mV] * 1.1 >= [Negative PD at 200 mV].

and - [Positive PD at 200 mV] * 0.9 <= [Negative PD at 200 mV].

THEN:

- Bonding material is missing or incompletely cured. Certainty 7. and- Exe (PHAFII).

A number is assigned to each rule: '3' in this example. Then, the keyword 'IF' indicates the beginning of the list of conditions, premises or antecedents of the rule. The first condition is true, if and only if, the user answer to the question: 'Insulation system is made of?' is 'mica folium' or 'asphalt-mica'. The second condition is true, if and only if, the variable [positive PD at 50 mV] multiply by 1.1 is greater or equal to the variable [Negative PD at 50 mV]. The same applies to the rest of the conditions. If one of these conditions happens to be false, because the user answer is different than expected, or because some variable value do not match the required condition, then the rule is false, and the inference machine of the expert system searches for another rule.

On the other hand, if all the conditions of a rule are true, then the rule is true and its conclusion is 'fired': 'Bonding material is missing or incompletely cured'. The word 'Certainty' at the end of a rule means 'the degree of certainty' or belief that the human expert has on the rule and it ranges from 0 to 10, where 10 means that the expert is absolutely certain of what the rule states.

With this production rule, the expert system can identify, with 70% certainty or reliability, that 'Bonding material is missing or incompletely cured' if the 'insulation system is made of mica folium or asphalt-mica', as stated in first condition, and if the 'positive partial discharge activity' is similar (within ten percent) to the 'negative partial discharge activity' at 'pulse magnitudes' of 50 mV and 200 mV, as stated in the rest conditions.

Then, our neural net paradigm is called using the command *Exe(PHAFII)*, where PHAF II is the module that handles the neural net. Algorithmic details of PHAF II are in [3]. We used the neural net to take advantage of the enormous amount of information currently available in many electric generator databases. Data from the partial discharge graphs are normalized within the range [0,1] and then fed to the PHAF II neural net. The neural net, previously trained with normalized data from graphs which are typical patterns of abnormal situations, performs the recognition of the fed graph and computes the percentage of similarity using three criteria:

- a) 10,000 base, where the graphs are compared using a lineal scale from 0 to 10,000 of frequency units. With this criterion, the differences or likenesses of the graphs have the same weight at high and low frequencies.
- b) Logarithmic base, where the graphs are compared using a logarithmic scale from 0 to 10,000 of frequency units. With this criterion, the differences or likenesses of the graphs are adjusted with more weight given to the differences in the low frequencies (0 to 100) and less weight to the differences at the high frequencies (100 to 10,000).
- c) Central base, where the graphs are compared using as a reference the pattern graph.

With the mean (average) of these three criteria, we obtain a final certainty factor. This factor indicates the similarity of the fed graph and the pattern graph. If the final factor is greater than 70% the system displays the screen shown in Fig. 1.

From Fig. 1, it is observed that the system displays the graph being recognized, the diagnosis, and eight certainty factors. Four of these correspond to a 'Global' analysis (GCF), where the certainty factor is computed as the mean of the likenesses or differences at all the points of the graphs. The other four certainty factors, called 'Local' (LCF), are obtained from the same point on the graphs where there exists the greatest distance between the graphs (the test graph and the pattern graph).

We planning, as future work, incorporate more human and data knowledge to this system. To efficient this phase we will investigate about automatic elicitation tools.

3 Flashover Forecasting

To approach the flashover on high-voltage insulators forecasting, we developed and integrated four data mining tools that combine the ID3 algorithm [4] and the nearest neighbor case-based reasoning method [5]. The first tool uses data mining to build a classification or decision tree from historic data, the second generates production rules, the third operates the decision tree as an expert system, and the last, makes tests with unseen cases to evaluate forecasting accuracy. The results were compared against other

classic machine learning tools like C4.5, FOIL, CN2 and OC1, and we obtain similar or better solutions.

To perform the experiments, data from an N-120P high-voltage insulator were registered during 21 days (504 examples of meteorological and surface resistance values). The attributes used were: hour of the day, wind direction, wind velocity, temperature, precipitation, dew temperature, barometric pressure, relative humidity and absolute humidity. The class attribute is the *surface resistance*, a variable correlated with the flashover phenomenon.

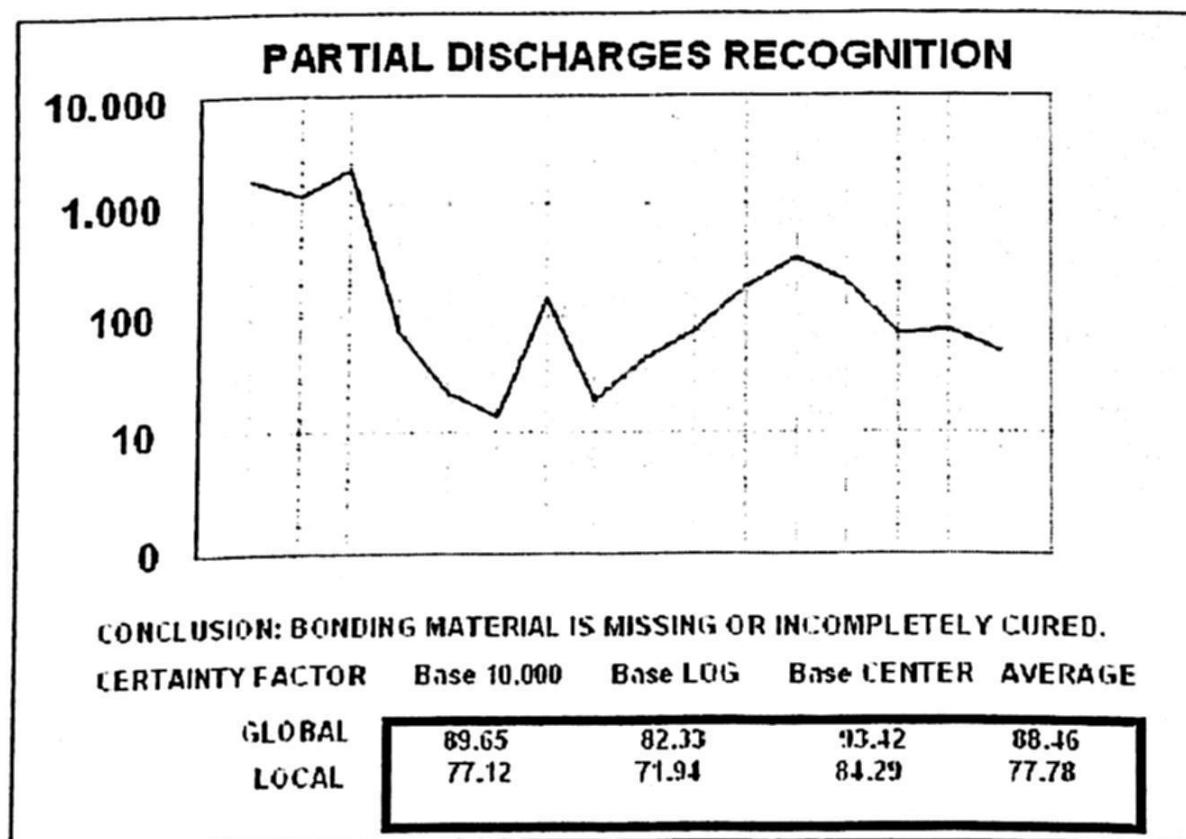


Fig.1 EGFD final display

Three classes were assigned to the surface resistance: "low" (for values between 1,013 and 3,489 kilo ohms), "medium" (for the interval 3,490/ 7,992) and "high" (for the interval 7,993/ 11,482).

With these tools, we could successfully determine:

- The relationship among the environmental variables and the surface resistance class.
- The variables that have more impact on the phenomenon, and the variables that hardly affect the surface resistance behavior.

- The circumstances that cause the surface resistance to be low.
- The surface resistance value 24 hours ahead, with an accuracy of 83%.

Table 1 shows comparison results. In the future we will work in the simplification and reduction of the production rule quantity generated for the proposed tool, to facilitate the interpretation of the discovered knowledge. We will work too with more data to try to forecast with more anticipation the flashover event.

Table 1. Results comparison.

TOOL	Proposed	C4.5	FOIL	CN2	OC1
TEST-Acc	82.9	82.9	62.9	82.4	80.1
TRAIN-Acc	99.6	85.3	96.8	92.0	81.5
# Rules	496	9	61	52	NA
Default	Unknown	Low	NA	High	NA
Time(secs)	3.68	143	10.7	74	643

where:

TEST-Acc = Accuracy to forecast unknown cases (%).

TRAIN-Acc = Accuracy to forecast the same examples used in learning stage (%).

Rules = Number of generated rules.

Default = Default class (if no rule applies).

Time (secs) = Seconds required to generate the results.

NA = Characteristic not available by the tool.

4 Electric Utilities Data Mining

This case is about extract knowledge from data. Some well-known data mining tools (C4.5, CN2, FOIL and PEBLS) were applied and evaluated for the task to obtain expert knowledge. This was done on a real power generation database with thermoelectric and hydroelectric Mexican electric utilities information over eight years of historic data. We evaluated accuracy, knowledge amount reduction and processing time. We analyzed the expert system rules (extracted knowledge) and we propose an architecture of an integrated knowledge discovery system for this electric power generation database [6].

For this research, personnel of the Performance Control and Informatics Unit of Federal Commission of Electricity in Mexico (CFE) selected the data. One table was built with 32 variables and 1,110 records corresponding to thermo and hydroelectric information for the years 1988 to 1995.

The 32 variables include the following: power plant identifier, date, plate and effective capacity; unavailability by type of failure; outage equivalent hours due to decrements,

number of outages and outage hours due to failure and due to routine and corrective maintenance and other causes; fuel kilocalories; net and gross generation; permanent workers used in maintenance, in operation, and in other activities; additional workers used in maintenance, in operation, and other activities; equivalent substitution workers in maintenance, in operation, and in other activities; total personnel positions; accidents that cause lost of time; accidents in transit; days lost due to accidents; days lost due accidents in transit; sum of disabilities in percent; and various expenses.

With this data set, a supervised data mining knowledge extraction was outlined. We used the variable Power Plant Factor (PF), as the "class" or focus of attention for the experiment. To calculate the PF the following formula was used:

$$PF = \frac{\text{Gross Generation}}{\text{HP} * \text{Net Generation}} * 100 \quad (1)$$

where HP (hours per period) is equal to 8,760 hours (365 days by 24 hours). It was found that only one rule describes the knowledge for 'excellent' plant factor for hydroelectric utilities, with 85.7% certainty:

IF:
 Unavailability due to failure (%) <= 9.375 and
 Unavailability due to maintenance (%) <= 0.520 and
 Unavailability due to other causes (%) <= 32.560 and
 Permanent Workers (Rest) > 822
 THEN: The Plant Factor is Excellent [certainty 85.7%]

Only one rule describes the knowledge for 'excellent' plant factor for thermoelectric utilities, with 92.3% certainty:

IF:
 Effective Capacity (MW) <= 298 and
 Gross Generation (GWH) > 1,721
 THEN: The Plant Factor is Excellent [certainty 92.3%]

We found that the variables that most affect the hydroelectric plant factor turned out to be: Unavailability due to other causes (63%), Gross generation (59%), and Effective capacity (48%). For thermoelectric utilities the variables were: Gross generation (87%), Effective capacity (83%), and Unavailability due to failure (48%). A summary of the results is shown in Table 2. From the experience obtained in the development of the experiments described, the need of having a system to facilitate the process of knowledge discovery using data mining algorithms and the exploration of various alternatives that improve the quality of the extracted knowledge (expert system rules) was evident.

So we proposed the creation of the following data mining modules:

- User Interface: allows the user to have an integrated environment, which shows the user a screen from which he can choose different options to accomplish the data mining and to obtain the results.

Table 2. Comparison of the number of Errors*

Plant Factor	C4.5	C4.5*	CN2	FOIL	FOIL*
Very-Low	4	7	8	1	1
Low	37	83	125	3	14
Regular	31	54	79	3	23
Average	23	67	87	10	7
Good	28	42	61	1	9
Very-Good	12	29	39	6	7
Excellent	5	22	23	0	0
Total	140\13.5%	304\29.2%	422\40.6%	24\2.3%	61\5.8%

Errors* = Number of cases misclassified using unseen data

Commentaries: FOIL has the better classification efficiency, followed by C4.5

C4.5 = results of the 'composite rule set'

C4.5* = results of the 'trial 0'

FOIL* = using similar attributes grouping

EXECUTION TIMES:

C4.5 = more than 30 mins.

CN2 = 10 mins.

FOIL = 128.1 secs.

FOIL* = 189.6 secs.

- Pre-Processing: this module handles different options to prepare the information of the database before the application of the mining algorithm. This module allows, among other things, the addition or deletion of columns and rows, clustering (using several methods like ChiMerge, 1R, Chi2, etc.) of continuously valued variables to group them in (a few) labeled classes, feature selection methods and to automatically prepare the data to the format required by the mining tool.
- Mining tools: the user selects from among several data mining tools, the one to be applied to the preprocessed data. Usually, it is necessary to try different algorithms due to the fact that there does not exist a perfect tool, but rather, depending on the data, some algorithms perform better than others.
- Post-Processing: through this module, the user may request the conversion of the extracted knowledge by the mining tool in a representation that it will be easier for him to understand; again, it does not exist "the best" representation of knowledge, since it depends on the user preferences. Some knowledge representations are: production rules, decision trees, graphics (OLAP), characteristic tables (prime relation tables and feature tables), Horn clauses, and prototypes.

These ideas were proposed by us before tools like Weka, Orange, Elvira, and others, arrived to the data mining community. However, still nowadays several issues related with the proposed modules are open for research, like data quality tools (profiling, cleansing, etc.), knowledge representation and visualization tools, etc.

5 Illicit Users Pattern Recognition

To process this problem, we have to realized feature selection pre-processing task due to the database size and because of noise data problems. The problem domain can be expressed thus: CFE faces the problem to accurately detect customers that illicitly use energy, and consequently, CFE tries to reduce the losses due to this concept. At present time, a lot of historical information is stored in the Commercial System (SICOM), an electric billing database. SICOM was created mainly to register the users contract information, and the invoicing and collection data; this database has several years of operation and has a great amount of accumulated data (millions of records).

To make feasible the mining of this large database, in an effective and efficient way, we firstly realized an evaluation of different filter-ranking methods for supervised learning. The evaluation took into account not only the classification quality and the processing time obtained after the filter application of each ranking method, but also it considered the discovered knowledge size, which, the smaller, the easier to interpret.

Also the boundary selection topic to determine which attributes must be considered relevant and which irrelevant was approached, since the ranking methods by themselves do not give this information. We proposed an extension, simple to apply, that allows unifying the criterion for the attributes boundary in the different evaluated ranking methods [7].

Based on the experimentation results, we proposed a heuristic that looks for the efficient combination of ranking methods with the effectiveness of the wrapper methods. Although our work focuses on the SICOM data, the lessons learned can be applied to other real world databases with similar problems.

Recently, to process this problem more efficiently and accurately, we proposed several competitive metrics and algorithms for feature selection considering inter-dependencies among nominal attributes (*buBF* method) [8] or numeric attributes (*dG* method) [9]. Some results and comparisons against other feature selection methods in Weka [10] and Elvira [12] tools are shown in Table 3.

Table 3. J4.8's accuracies for 10-fold-cross validation using the features selected by each method (Electric billing database).

Method	Total features selected	Accuracy (%)	Pre-processing time
CFS	1	90.18	9 secs.
<i>dG</i>	2	90.70	43 secs.
<i>vG</i>	3	94.02	0.7 secs.
Bhattacharyya	3	90.21	6 secs.
Matusita distance	3	90.21	5 secs.
ReliefF	4	93.89	14.3 mins.
Euclidean distance	4	93.89	5 secs.
Kullback-Leibler 1	4	90.10	6 secs.
Mutual Information	4	90.10	4 secs.
<i>buBF</i>	5	97.50	1.5 secs.
Kullback-Leibler 2	9	97.50	6 secs.
OneR	9	95.95	41 secs.
Shannon entropy	18	93.71	4 secs.
ChiSquared	20	97.18	9 secs.
All attributes	24	97.25	0

Furthermore, these ideas was applied successfully to other well known databases [12], as Table 4 shows. So we can conclude that the proposed metrics and feature selection methods are valuable tools to detect relevant attributes.

In the near future we will work in developed feature selection methods for mixed data, this is to say, for nominal and numeric attributes at the same time.

6 Conclusions and Future Work

We have presented four data mining applications in the Mexican power industry, and the way that we approached each one of them. From the experimentations presented we think that our proposed methods represents promising alternatives, compared to other methods, because of its acceptable performance. At Table 5 we resume our experiences: we outline advantages, drawbacks and possible solutions that we think it have to be done in the near future to maximizing the usefulness of the data mining techniques that we used in our works.

Table 4. J4.8's accuracies using the features selected by each method for five UCI datasets.

Method	Autos (25/205/7)			Horse-c (27/368/2)			Hypothyroid (29/3772/4)			Sonar (60/208/2)			Ionosphere (34/351/2)			Avg. Acc
	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	
All atts	25	82	0	27	66	0	29	99	0	60	74	0	34	91	0	82.4
<i>buBF</i>	9	77	0.2	4	72	0.22	5	97	0.31	10	74	0.6	4	90	0.9	82.0
<i>vG</i>	8	75	0.01	3	69	0.02	4	95	0.2	11	73	0.03	4	91	0.3	80.6
<i>dG</i>	7	75	12	2	68	14	5	95	26	9	75	14	3	88	18	80.2
CFS	6	74	0.05	2	66	0.04	2	96	0.3	18	74	0.09	8	90	3	80.0
ReliefF	11	74	0.4	3	66	0.9	6	93	95	4	70	0.9	6	93	4	79.2
SOAP	3	73	0.01	3	66	0.02	2	95	0.2	3	70	0.02	31	90	0.01	78.8
Mutual I	3	72	0.9	4	68	1	2	90	1.4	18	73	1	3	86	1	77.8
OneR	5	70	0.8	3	67	1	3	88	1.3	12	72	1	4	85	1	76.4
KL-1	3	71	0.9	4	61	1.2	3	92	1.7	16	70	1	2	86	1	76.0
KL-2	4	68	0.9	4	62	1.1	2	89	1.5	11	68	1	3	83	1	74.0
Matusita	3	66	1.7	3	61	2.3	2	91	3.3	17	68	2.5	2	83	2	73.8
Bhattach	3	67	0.8	3	60	1	1	90	1.4	9	68	1	2	83	1	73.6
Euclidean	2	66	1	3	62	1.4	2	90	1.2	10	67	1.1	2	82	1	73.4
ChiSqua	3	67	1	2	60	1.6	3	88	1.3	11	65	1.2	2	80	1	72.0
Shannon	4	66	0.9	4	61	1.3	2	87	1.6	9	66	1	2	80	1	72.0

"(25/205/7)" means (attributes/ instances/ classes) for Autos dataset, and so on.

TF=Total features selected Ac=Accuracy (%) Pt=Pre-processing time (secs.)

In our opinion, a great variety of Mexican power industry applications are still waiting to be tackled with data mining techniques, but we need develop more and sophisticated tools to accomplish the challenges. Some future work includes problems with real and very large power system databases such as the national power generation performance database, the national transmission energy control databases, the de-regulated energy market database, and the Mexican electric energy distribution database. Also, we need apply statistical tests to observe if the differences in accuracies, processing time or another parameter of the proposed methods are really significant.

Table 5. Our experiences and recommendations.

Approach used	Advantage	Drawback	Possible solution
Expert System	Representation of human-expert knowledge in a natural way.	Complex elicitation process.	Develop more sophisticated and computer aided elicitation tools.
Neural Network	Captures knowledge from numeric data.	It needs manual tuning. Discovered knowledge is in a black box.	Develop tools for dynamical tuning and to extract knowledge from neural inter-connections.
Induction Tree	Captures and shows knowledge from nominal data in an explicit way.	It needs previous data discretization. Obtained results are not very precise.	Develop tools for automatic and efficient data discretization. Improve output thru post-processing-visualization tools.
Data Mining	Discovers and shows hidden knowledge from data.	It needs an integration of the pre-processing, processing and post-processing phases.	Construct a integrated system with: data quality process, final user easy of interpret knowledge representation and visualization tools.
Feature Selection	Detects relevant attributes and reduces problem size.	There are no infallible method.	Research for metrics that evaluate attribute relevance (numeric and nominal data at once) in an effective way.

References

1. The 24th Annual International Conference on Machine Learning (ICML-2007) June 20-24, Oregon State University, USA, 2007.
2. Piatetsky-Shapiro, G. et al, Knowledge Discovery in Databases: An Overview, In Knowledge Discovery in Databases, Piatetsky-Shapiro, G. eds., Cambridge, MA, AAAI/MIT, 1991, pp 1-27.
3. Mejía, M., Rodríguez, G., A New Neural Network Paradigm for Power Systems Applications, Proceedings of the IASTED International Conference on Power Systems and Engineering, Vancouver, Canada 1992, pp. 41-48.
4. Quinlan, J., Discovering Rules by Induction from Large Collections of Examples, Expert Systems in the Micro-Electronic Age, Michie, D., (ed), Edinburgo, Escocia, Edinburgh University Press, 1979.

5. Mejía, M., Rodríguez, G., Montoya, G., Knowledge discovery in high-voltage insulators data, *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Proceedings of the Tenth International Conference, Atlanta, Georgia, USA, June 1997, pp. 223-230.
6. Mejía, M., Rodríguez, G., Obtaining expert systems rules using data mining tools from a power generation database, *Expert Systems with Applications*, J.Liebowitz (ed), 14(1/2) Pergamon, 1998, pp. 37-42.
7. Mejía, M., Rodríguez, G., Arroyo, G., Morales, Feature selection-ranking methods in a very large electric database. *MICAI 2004: Advances in Artificial Intelligence, 3rd Mexican Int. Conf. on Artificial Intelligence*, Springer Berlin, April, pp. 292-301.
8. Mejía, M., Morales, E. 2006. Feature Selection in an Electric Billing Database Considering Attribute Inter-dependencies. In Petra Perner (ed) *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining: 6th Industrial Conference on Data Mining*, ISBN: 3-540-36036-0, ISSN: 0302-9743, LNCS 4065, Springer Berlin / Heidelberg, Leipzig, Germany, pp. 284-296.
9. Mejía, M., Morales, E. 2007. Two Two Simple and Effective Feature Selection Methods for Continuous Attributes with Discrete Multi-Class. *MICAI 2007 6th Mexican Int. Conf. on Artificial Intelligence*, LNAI 4827, Springer Berlin, November, pp. 452-461.
10. www.cs.waikato.ac.nz/ml/wcka, 2004.
11. www.ia.uned.es/~elvira/, 2004.
12. Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mlearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

Using WEKA for Semantic Classification of Spanish Verb-Noun Collocations

Olga Kolesnikova and Alexander Gelbukh

Center for Computing Research, National Polytechnic Institute
Mexico City, 07738, Mexico
kolesolga@gmail.com, www.gelbukh.com

Abstract. Collocations, or restricted lexical co-occurrence, can be classified according to their semantics represented by the formalism of lexical functions in the frame of Meaning-Text Theory. We use learning algorithms implemented in WEKA to classify Spanish verb-noun collocations according to lexical functions. Experiments were made on a manually built corpus of verb-noun pairs. WEKA classifiers were tested for detection of two lexical functions, namely Oper1 and CausFunc0. Some WEKA classifiers show better performance than state-of-the-art results.

1 Introduction

Knowledge of collocation is important for natural language processing because collocation comprises the restrictions on how words can be used together. Automatic extraction techniques produce a list of collocations. Such lists are more valuable if collocational data is tagged with semantic information. In this section, we will consider the notion of collocation, approaches to classification of collocations, and lexical functions as semantic typology for collocational classification.

1.1 Collocations

Collocation has been a controversial issue in linguistic research for a number of decades. Many attempts have been made to define collocation. These definitions are based on various criteria: statistical, lexical, functional, structural, semantic. We make use of a semantic definition of collocation given in [9]. This definition affirms that collocation is a combination of two lexical items in which the semantics of one of the lexical items (called the “base” of collocation) is autonomous from the combination it appears in, and where the other lexical item (called the “collocate”) adds semantic features to the semantics of the base. In other words, collocation is a binary word combination in which the base is used in its “normal” meaning and the collocate is used in a “non-typical” meaning. As an example, consider a Spanish collocation *dar un paseo* ‘give a walk’. The noun *paseo* is the base and is used in its typical meaning, but the verb *dar* here means ‘do’ although its most frequent meaning is ‘give’. There are other verb-noun collocations where the verbal collocate acquires the same

meaning 'do': *cometer un error* ('make a mistake'), *ejercer control* 'exercise control', *hacer una pregunta* 'ask a question', *realizar el esfuerzo* 'make an effort'. We can observe that all these collocations can be characterized by the meaning pattern 'do something'. The next group of collocations reveals another meaning pattern – 'cause something to exist': *causar daño* 'cause damage', *abrir una posibilidad* 'open a possibility', *establecer una regla* 'establish a rule', *formar un grupo* 'form a group'. Therefore, semantic patterns 'do something', 'cause something to exist' represent semantic contents of respective collocations.

1.2 Classification of collocations

Collocations are classified on the basis of various criteria. According to their grammatical structure, collocations can be grammatical and lexical [5]. Depending on the part of speech of the base and the collocate, collocations can be adjective-noun, verb-noun, adverb-adjective, verb-adverb, etc. [3]. If frequency of collocational parts in corpus is applied as a classification principle, then 'upward' and 'downward' collocation are distinguished by Sinclair [13]. Wanner proposed a semantic classification of collocation based on the taxonomy of lexical functions [17].

1.3 Lexical function

Lexical function (LF) is a formalism developed within the Meaning-Text Theory [8] [10] to represent semantic and syntactic structure of collocation. It has a general form

$$LF_{n_1...n_k}(b) = c,$$

where b is the base of a collocation, c is the collocate. In terms of the Meaning-Text Theory, b is called the keyword, and c – the LF value. "LF" in the formula stands for the name of lexical function. The LF name is an abbreviated Latin word which denotes the semantic contents of collocations. The string of positive integers " $n_1...n_k$ " is optional. The integer value indicates semantic valency of the keyword. The position of the integer in the string represents the syntactic function of the word used to fill in the corresponding semantic valency (first position signifies the subject, second – direct object, etc). Let us consider a few examples.

$Oper1(paseo) = dar$. The keyword (the base of the collocation) is *paseo*, 'walk'. The LF value is *dar*, 'give'. The lexical function has the name "Oper" from Latin *operari* – 'do, carry out'. Since *paseo* denotes an action, it can be viewed in its verbal aspect. In this case, its first semantic role is the agent. The integer "1" means that the word used to lexicalize the role of agent, functions as a subject in a sentence, for example *El doctor salió a dar un paseo por la tarde*. 'The doctor left to take a walk in the afternoon', where the doctor is the agent.

$Func0(posibilidad) = existir$. "Func" is from Latin *functionare* – to 'function', so respective collocations have meaning 'something functions, happens, takes place'. The integer "0" means that no semantic role is lexicalized as subject, but the keyword itself functions as the subject: *No existe la posibilidad de realizar este proyecto*. 'No possibility exists to realize this project'.

Manifl(*problema*) = *plantear*. "Manif" is from Latin *manifestare* – to 'manifest'. Respective collocations mean that the agent of the verb reveals something that it becomes apparent. *Plantear problema* in Spanish corresponds to *pose a problem* in English.

The above examples demonstrate so-called simple lexical functions which formalize a single semantic element, or one meaning like 'do', 'function', 'manifest'. However, there are many cases when collocations have more complex meaning which is formed by a combination of two or more "single" meanings. This phenomenon is captured by complex LFs. Before giving examples of complex LF, it should be mentioned that there exist simple LFs that are seldom used independently but more often constitute parts of complex LFs. Table 1 lists names of such simple LFs and their respective meanings taken from [10].

Table 1. Examples of simple LFs used in complex LFs more often than independently.

LF	Meaning	Comment
Incep	Lat. <i>incipere</i> – 'begin'	something begins occurring
Cont	Lat. <i>continuare</i> – 'continue'	something continues occurring
Fin	Lat. <i>finire</i> – 'cease'	something ceases occurring
Caus	Lat. <i>causare</i> – 'cause'	do something so that a situation begins occurring
Liqu	Lat. <i>liquidare</i> – 'liquidate'	do something so that a situation stops occurring

A complex LF is a combination of two or more simple LFs. Table 2 presents a few complex LFs and their meanings, for each complex LF examples of collocations are presented. "K" in the column "Meaning" stands for the keyword.

Table 2. Examples of complex LFs.

LF	Meaning	Keyword	LF Value	Collocations
IncepOper1	begin to do K	<i>proceso</i> 'process' <i>responsabilidad</i> 'responsibility'	<i>iniciar</i> 'begin' <i>asumir</i> 'assume'	<i>iniciar el proceso</i> 'begin the process' <i>asumir la responsabilidad</i> 'assume the responsibility'
ContOper1	continue to do K	<i>contacto</i> 'contact' <i>camino</i> 'road'	<i>mantener</i> 'maintain' <i>seguir</i> 'follow'	<i>mantener el contacto</i> 'maintain the contact' <i>seguir el camino</i> 'follow the road'
CausFunc0	cause that K comes into existence	<i>efecto</i> 'effect' <i>explicación</i> 'explanation'	<i>producir</i> 'produce' <i>dar</i> 'give'	<i>producir el efecto</i> 'produce the effect' <i>dar una explicación</i> 'give an explanation'
LiquFunc0	do something that K ceases to exist	<i>vida</i> 'life' <i>problema</i> 'problem'	<i>quitar</i> 'take away' <i>evitar</i> 'avoid'	<i>quitar la vida</i> 'take (one's) life' <i>evitar el problema</i> 'avoid the problem'

As mentioned before, collocations are classified on the basis of various criteria. Since LFs represent semantic patterns of collocations, LF taxonomy can be used to build a semantic classification of collocations. Besides, the taxonomy of LFs is advantageous because it groups collocations according to language-independent generalized semantics and characteristic syntactic patterns. Implemented in a computer readable dictionary of collocations, a classification by lexical functions will

allow effective use of collocations in natural language applications including parsers, high quality machine translation, systems of paraphrasing and computer-aided learning of lexica [2].

The rest of the paper is organized as follows. Section 2 summarizes previous research on automatic detection of LFs. Section 3 defines the objective of this work. We discuss the experimental results in Section 4. Section 5 presents conclusions and outlines future work.

2 Related work

2.1 Automatic Detection of Lexical Functions

There have been made a few attempts to detect LFs automatically. Wanner approached automatic detection of LFs as the task of automatic classification of collocations according to LF typology [17]. He applied nearest neighbor machine learning technique to classify Spanish verb-noun pairs according to nine LFs selected for the experiments. The distance of candidate instances to instances in the training set was evaluated using path length in hyperonym hierarchy of the Spanish part of EuroWordNet [16]. An average f-score of about 70% was achieved in these experiments. The largest training set included 38 verb-noun pairs (for LF CausFunc0) and all test sets had the size of 15 instances.

Alonso Ramos *et al.* [1] propose an algorithm for extracting collocations following the pattern "support verb + object" from FrameNet corpus of examples [12] and checking if they are of the type *Opern*. This work takes advantage of syntactic, semantic and collocation annotations in the FrameNet corpus, since some annotations can serve as indicators of a particular LF. The authors tested the proposed algorithm on a set of 208 instances. The algorithm showed accuracy of 76%. Alonso Ramos *et al.* conclude that extraction and semantic classification of collocations is feasible with semantically annotated corpora. This statement sounds logical because the formalism of lexical function captures the correspondence between the semantic valency of the keyword and the syntactic structure of utterances where the keyword is used in a collocation together with the value of the respective LF.

2.2 Collocation classification according to LF

Wanner *et al.* [18] experiment with the same type of lexical data as [17], i.e. verb-noun pairs. The task is to answer the question: what kind of collocational features are fundamental for human distinguishing among collocational types. The authors view collocational types as LFs, i.e. a particular LF represents a certain type of collocations. Three hypotheses are put forward as possible solutions, and to model every solution, an appropriate machine learning technique is selected. Below we list the three hypotheses and the selected machine learning techniques.

1. Collocations can be recognized by their similarity to the prototypical sample of each collocational type; this strategy is modeled by the Nearest Neighbor technique.
2. Collocations can be recognized by similarity of semantic features of their elements (i.e., base and collocate) to semantic features of elements of the collocations known to belong to a specific LF; this method is modeled by Naïve Bayesian network and a decision tree classification technique based on the ID3-algorithm.
3. Collocations can be recognized by correlation between semantic features of collocational elements; this approach is modeled by Tree-Augmented Network Classification technique.

In classification experiments, the authors deal with two groups of verb-noun collocations. The first group includes only verb-noun pairs where nouns belong to the semantic field of emotions. In the second group, nouns are field-independent. We will compare the results for the second group of collocations with WEKA performance in Section 4.3. It should be mentioned also, that having proposed three hypotheses, the authors have not yet demonstrated their validity by comparing the performance of many machine learning techniques known today, but apply only four learning algorithms to illustrate that three human strategies mentioned above are practical. This will be considered in more detail in Section 4.3.

3 Objective

The aim of this paper is to test WEKA methods for classification of collocations according to LFs. We apply WEKA classifiers to a corpus of Spanish verb-noun combinations and train the system to detect two lexical functions chosen for the experiments. For each LF in question, the classifier with best results is identified. The obtained results are compared with those in [18] for verb-noun collocations with field-independent nouns.

4 Experimental results

4.1 Experimental Methodology

We apply machine learning techniques as implemented in the WEKA version 3-6-2 learning and data mining toolset [6] [21]. The data as described in Section 4.2 was supplied to 67 classifiers of various classes. We evaluated the performance of WEKA classifiers by comparing precision, recall, and f -measure (weighted average values over all classes). The precision is the proportion of the examples which truly have class x among all those which were classified as class x . The recall is the proportion of examples which were classified as class x , among all examples which truly have class x . The f -measure is $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$ [19].

4.2. Data

For compiling data sets, we used a list of verb-object pairs extracted by the Word Sketch Engine [7] from Spanish web corpus [14] and ranked by frequency. 1000 most frequent pairs were annotated with part of speech, Spanish WordNet version 200611 [16] [15] word senses, and LFs by human experts.

Table 3 gives the number of samples in the list for all LFs found in this list. The number of instances that were annotated as free word combinations was 198; 61 pairs were qualified as errors (for example, some pairs contained symbols like “, ©, -- instead of words, other pairs turned out to be combinations “verb + past participle”).

Table 3. LFs with the respective number of samples for 1000 most frequent verb-noun pairs in Spanish Web Corpus.

LF	Number of samples	LF	Number of samples
Oper1	157	Real2	3
CausFunc0	102	IncepFunc0	3
CausFunc1	60	MinusCausFunc0	3
Real1	45	ManifCausFunc0	2
Func0	22	LiquFunc0	2
Oper2	16	AntiReal3	1
IncepOper1	14	Oper3	1
ContOper1	11	PlusCausFunc0	1
Copul	9	FinOper1	1
Manif	9	MinusCausFunc1	1
Caus2Func1	9	FinFunc0	1
PlusCausFunc0	6	PermOper1	1
PlusCausFunc1	5	Real3	1
PerfOper1	4	Caus1Oper1	1
Caus1Func1	3	PerfFunc0	1

Since the most frequent LFs in our data were Oper1 and CausFunc0, we constructed one data set for Oper1 and another set for CausFunc0. These are the steps we followed to build a training set for each LF out of the source file which contained the list of 1000 most frequent pairs annotated with POS, word senses, and LFs as described above.

1. Samples of the LF in question are marked as positive examples and are included in the training set.
2. Verb-noun pairs of all other LFs are marked as negative examples and are included in the training set. Pairs qualified as errors are not made a part of the training set. Free word combinations are considered a lexical function and are incorporated into the training set as negative examples.
3. All hyperonyms for every word in the set of positive and negative examples are extracted from the Spanish WordNet. Synsets to which the words in the examples belong are considered zero-level hyperonyms.
4. The training set for WEKA tool has the Attribute-Relation File Format (ARFF) [22]. Every hyperonym in the set of Step 3 is considered as a nominal attribute which can take one of two values: “1” if it is a hyperonym of any word in a given

Table 5. Performance of WEKA Classifiers of classes bayes, function, meta, and rules on Oper1 and CausFunc0 data sets. P stands for precision, R – for recall, F – for *f*-measure.

Classifier class	Classifier	Oper1			CausFunc0		
		P	R	F	P	R	F
bayes	AODE	0.841	0.845	0.842	0.883	0.89	0.856
	AODEsr	0.802	0.808	0.803	0.852	0.83	0.84
	BayesianLogisticRegression	0.927	0.927	0.927	0.901	0.907	0.903
	BayesNet	0.836	0.832	0.834	0.88	0.883	0.881
	HNB	0.857	0.859	0.852	0.845	0.877	0.837
	NaiveBayes	0.837	0.84	0.838	0.861	0.885	0.858
	NaiveBayesSimple	0.837	0.84	0.838	0.861	0.885	0.858
	NaiveBayesUpdateable	0.837	0.84	0.838	0.861	0.885	0.858
	WAODE	0.838	0.84	0.839	0.883	0.897	0.881
functions	LibSVM	0.507	0.712	0.593	0.765	0.875	0.816
	Logistic	0.913	0.905	0.907	0.9	0.89	0.894
	RBFNetwork	0.828	0.833	0.828	0.854	0.88	0.842
	SimpleLogistic	0.92	0.92	0.92	0.9	0.907	0.902
	SMO	0.922	0.922	0.922	0.911	0.915	0.913
	VotedPerceptron	0.894	0.896	0.894	0.88	0.893	0.883
	Winnow	0.677	0.687	0.681	0.812	0.764	0.785
meta	AdaBoostM1	0.828	0.808	0.777	0.829	0.875	0.819
	AttributeSelectedClassifier	0.913	0.913	0.913	0.912	0.919	0.913
	Bagging	0.913	0.913	0.913	0.917	0.922	0.918
	ClassificationViaClustering	0.599	0.666	0.616	0.772	0.774	0.773
	ClassificationViaRegression	0.894	0.893	0.894	0.923	0.923	0.923
	CVParameterSelection	0.507	0.712	0.593	0.765	0.875	0.816
	Dagging	0.89	0.889	0.889	0.895	0.905	0.887
	Decorate	0.891	0.887	0.889	0.896	0.899	0.898
	END	0.91	0.91	0.91	0.909	0.916	0.91
	EnsembleSelection	0.917	0.917	0.917	0.926	0.929	0.927
	FilteredClassifier	0.91	0.91	0.91	0.909	0.916	0.91
	Grading	0.507	0.712	0.593	0.765	0.875	0.816
	LogitBoost	0.914	0.915	0.914	0.898	0.907	0.899
	MultiBoostAB	0.819	0.769	0.709	0.765	0.875	0.816
	MultiClassClassifier	0.913	0.905	0.907	0.9	0.89	0.894
	MultiScheme	0.507	0.712	0.593	0.765	0.875	0.816
	OrdinalClassClassifier	0.91	0.91	0.91	0.909	0.916	0.91
	RacedIncrementalLogitBoost	0.507	0.712	0.593	0.765	0.875	0.816
	RandomCommittee	0.874	0.868	0.87	0.895	0.902	0.898
	RandomSubSpace	0.879	0.875	0.867	0.884	0.897	0.877
	RotationForest	0.904	0.905	0.904	0.908	0.915	0.909
	Stacking	0.507	0.712	0.593	0.765	0.875	0.816
	StackingC	0.507	0.712	0.593	0.765	0.875	0.816
	ThresholdSelector	0.916	0.915	0.915	0.887	0.895	0.89
	Vote	0.507	0.712	0.593	0.765	0.875	0.816
rules	ConjunctiveRule	0.784	0.762	0.705	0.765	0.875	0.816
	DecisionTable	0.905	0.906	0.905	0.902	0.907	0.904
	JRip	0.914	0.915	0.914	0.932	0.933	0.932
	NNge	0.894	0.892	0.893	0.888	0.895	0.891
	OneR	0.819	0.769	0.709	0.877	0.893	0.871
	PART	0.911	0.912	0.911	0.896	0.899	0.897
	Prism	0.881	0.874	0.876	0.912	0.909	0.911
	Ridor	0.895	0.896	0.896	0.909	0.915	0.911
	ZeroR	0.507	0.712	0.593	0.765	0.875	0.816

Table 6. Performance of WEKA Classifiers of classes misc and trees on Oper1 and CausFunc0 data sets, P stands for precision, R – for recall, F – for *f*-measure.

Classifier class	Classifier	Oper1			CausFunc0		
		P	R	F	P	R	F
misc	HyperPipes	0.785	0.769	0.775	0.865	0.838	0.849
	VFI	0.846	0.849	0.847	0.875	0.856	0.864
trees	ADTree	0.916	0.916	0.916	0.913	0.919	0.915
	BFTree	0.917	0.917	0.917	0.921	0.925	0.922
	DecisionStump	0.819	0.769	0.709	0.765	0.875	0.816
	FT	0.92	0.92	0.92	0.91	0.912	0.911
	Id3	0.926	0.926	0.926	0.905	0.905	0.905
	J48	0.91	0.91	0.91	0.909	0.916	0.91
	J48graft	0.907	0.907	0.907	0.897	0.907	0.895
	LADTree	0.921	0.922	0.921	0.919	0.919	0.919
	RandomForest	0.87	0.863	0.865	0.897	0.906	0.899
	Random Tree	0.819	0.813	0.816	0.865	0.876	0.87
	REPTree	0.903	0.903	0.903	0.925	0.929	0.927
	SimpleCart	0.921	0.922	0.921	0.921	0.925	0.922

Table 7 presents the highest state-of-the-art results [18] for detection of Oper1 and CausFunc0. NN signifies the Nearest Neighbor technique, NB – Naïve Bayesian network, ID3 – a decision tree classification technique based on the ID3-algorithm, TAN – Tree-Augmented Network Classification technique. Experiments on ID3-algorithm were not done for CausFunc0.

Table 7. State-of-the-art results [18] for Oper1 and CausFunc0, P is precision, R – recall, F – *f*-measure.

LF	NN		NB		ID3		TAN	
	P	R	P	R	P	R	P	R
Oper1	0.65	0.55	0.87	0.64	0.52	0.51	0.75	0.49
CausFunc0	0.59	0.79	0.44	0.89	--	--	0.45	0.57

The best state-of-the-art result (recall 0.89) is achieved for CausFunc0 by applying Naïve Bayesian network. The highest result we obtained (recall 0.933) is for CausFunc0 by applying rules.JRip classifier. (Names of classifiers are given in the format <className>.<classifierName>). In both cases, results for CausFunc0 are higher than for Oper1. It means that instances of Oper1 have more similarity with the rest of examples in the list of verb-noun pairs than instances of CausFunc0.

The training set for CausFunc0 in [18] contained 53 positive examples and for Oper1 – 84 positive examples. In our experiments, the training set for CausFunc0 had 102 positive examples, and for Oper1 – 157 positive examples. Larger data sets improve the performance of classifiers and the obtained results are more statistically reliable.

A difference between data representation in our experiments and data sets used in [18] should be noted here. In [18], every word in the training set was accompanied by its synonyms and hyperonyms, its own Base Concepts (BC) and the BCs of its hyperonyms, its own Top Concepts (TC) and the TCs of its hyperonyms taken from the Spanish part of the EuroWordNet. We included only hyperonyms in the training

sets. Though WEKA classifiers were fed with less information in our case, it seems quite sufficient to produce better performance than in [18]. This phenomenon may remind us of the original intent of WordNet compilers who suggested to describe the meaning of any word by semantic relations only [11], like "is-a-kind-of" semantic relation of hyperonym hierarchy. Later, WordNet authors admitted that their previous assumption had been wrong and glosses were added to distinguish synonym sets. Though practical significance of glosses is generally accepted, we have seen that classifier accuracy is no worse if only hyperonyms are taken into account. Further research is needed to investigate how additional information, like that of semantic ontologies, changes classifier performance.

In the light of our results, let us consider the three methods of human recognition of collocations proposed in [18] and considered in Section 2.2.

Method 1. Collocations can be recognized by their similarity to the prototypical sample of each collocational type; this was modeled by Nearest Neighbor technique. Weka implements the nearest neighbor method in the following classifiers: rules.NNge, lazy.IB1, lazy.IBk and lazy.KStar [20]. Among these four classifiers, good results are obtained by rules.NNge for both Oper1 and CausFunc0. It demonstrates that Method 1 is feasible though does not produce very high quality results.

Method 2. Collocations can be recognized by similarity of semantic features of collocational elements to semantic features of elements of collocations known to belong to a specific LF; this was modeled by Naïve Bayesian network and a decision tree classification technique based on the ID3-algorithm. We tested three WEKA Naïve Bayesian classifiers – bayes.NaiveBayes, bayes.NaiveBayesSimple, bayes.NaiveBayesUpdateable [20]. All three classifiers show equal results, and the results for CausFunc0 are higher than for Oper1. ID3 algorithm is implemented in trees.Id3. This classifier gives better results than Naive Bayes and rules.NNge in Method 1.

Method 3. The third method was modeled by Tree-Augmented Network (TAN) Classification technique. As it is seen from Table 7, nearest neighbor algorithm gives better results in terms of recall than TAN. We did not apply TAN method in our experiments.

As it was mentioned before, the highest result obtained (recall 0.933) is produced by rules.JRip classifier. JRip classifier implements the RIPPER propositional rule learner [4]. The learning model is developed by iteration over a training subset, and by doing structure optimization to minimize error rate. More details can be found in [20].

5 Conclusions and future work

Our experiments have shown that verb-noun collocations can be classified according to semantic taxonomy of lexical functions using WEKA learning toolset. The best performance was demonstrated by rules.JRip classifier for lexical function CausFunc0. The highest result for detecting the lexical function Oper1 is given by bayes.BayesianLogisticRegression classifier. Both classifiers can be applied for high

quality semantic annotation of verb-noun collocations based on the taxonomy of lexical functions. This was demonstrated on Spanish material.

As future work, we plan to experiment with different ratios of training and test sets as well as experiment on English verb-noun collocations. We will evaluate the performance of WEKA classifiers for more lexical functions and analyze errors of classifiers.

We have seen that rules.JRip accuracy is high when a verb-noun collocation is represented as a set of all hyperonyms of the noun and all hyperonyms of the verb. We plan to explore how performance of classifiers changes if we add other data like word glosses to the training set.

Acknowledgements. We are grateful to Adam Kilgarriff and Vojtech Kovár for providing us a list of most frequent verb-noun pairs from the Spanish Web Corpus of the Sketch Engine, www.sketchengine.co.uk.

The work was done under partial support of Mexican Government: SNI, COFAA-IPN, PIFI-IPN, CONACYT grant 50206-H, and SIP-IPN grant 20100773.

References

1. Alonso Ramos, M., Rambow O., Wanner L.: Using semantically annotated corpora to build collocation resources. *Proceedings of LREC, Marrakesh, Morocco*, pp. 1154--1158 (2008).
2. Apresjan, Ju.D., Boguslavsky, I. M., Iomdin, L. L., Tsinman, L. L.: Lexical Functions in NLP: Possible Uses. In: Klenner, M., Visser, H. (eds) *Proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg*, 21-22 July 2000, pp. 55--72. Frankfurt am Main (2002)
3. Benson, M., Benson, E., Ilson, R.: *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamin Publishing Company (1997)
4. Cohen, W. W.: Fast effective rule induction. In: Prieditis, A., Russell S. (eds) *Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, CA, pp.115--123. Morgan Kaufmann, San Francisco (1995)
5. Gitsaki, C.: *The Development of ESL Collocational Knowledge*, Ph.D. thesis, Center for Language Teaching and Research, The University of Queensland, Brisbane, Australia (1996)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten I. H.: *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1 (2009)
7. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: *The Sketch Engine*. In *Proceedings of EURALEX 2004*, pp. 105--116 (2004)
8. Mel'cuk, I. A.: *Opyt teorii lingvisticskix modelej "Smysl? Tekst"* ('A Theory of the Meaning-Text Type Linguistic Models'). Nauka, Moscow (1974)
9. Mel'cuk, I. A.: *Phrasemes in Language and Phraseology in Linguistics*. In: Everaert, M., Van der Linden, E.-J., Schenk, A., Schreuder, R. (eds) *Idioms: Structural and Psychological Perspectives*, pp. 167--232. Lawrence Erlbaum, Hillsdale, NJ (1995)

10. Mel'cuk, I. A.: Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In: Wanner, L. (ed) *Lexical Functions in Lexicography and Natural Language Processing*, pp. 37--102. Benjamins Academic Publishers, Amsterdam, Philadelphia, PA (1996)
11. Miller, G.A: Foreword. In: Fellbaum, C. (ed.) *WordNet. An Electronic Lexical Database*, pp. xv--xxii. MIT Press, Cambridge, Mass. (1998)
12. Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. R., Scheffczyk, J.: *FrameNet II: Extended Theory and Practice*, <http://framenet.icsi.berkeley.edu/book/book.pdf>. ICSI Berkeley (2006)
13. Sinclair, J.: *Corpus Concordance Collocation*. OUP, Oxford (1991)
14. Spanish Web Corpus in SketchEngine, <http://trac.sketchengine.co.uk/wiki/Corpora/SpanishWebCorpus>
15. Spanish WordNet, http://www.lsi.upc.edu/~nlp/web/index.php?Itemid=57&id=31&option=com_content&task=view, last viewed March 26, 2010
16. Vossen P. (ed): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht (1998)
17. Wanner, L.: Towards automatic fine-grained classification of verb-noun collocations. *Natural Language Engineering*, vol. 10(2), pp. 95--143. Cambridge University Press, Cambridge (2004)
18. Wanner, L., Bohnet, B., Giereth, M.: What is beyond Collocations? Insights from Machine Learning Experiments. *EURALEX* (2006)
19. WEKA Manual for Version 3-6-2, <http://iweb.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3-6-2.pdf>
20. Witten, I. H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*, 2nd Edition. Morgan Kaufmann, San Francisco (2005)
21. The University of Waikato Computer Science Department Machine Learning Group, WEKA download, http://www.cs.waikato.ac.nz/~ml/weka/index_downloading.html, last viewed March 26, 2010
22. The University of Waikato Computer Science Department Machine Learning Group, Attribute-Relation File Format, <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>, last viewed March 26, 2010

Experimenting with Maximal Frequent Sequences for Multi-Document Summarization*

Yulia Ledeneva, René Arnulfo García-Hernández,
Anabel Vazquez-Ferreira, Nayely Osorio de Jesús

Autonomous University of the State of Mexico
Santiago Tianguistenco, México
ynledeneva@uaemex.mx
renearnulfo@hotmail.com

Abstract. In this work we address the task of multi-document text summarization which consists in reducing the contents of a collection of documents to a single short text, so that the user can judge about the contents of the whole collection upon reading only this short text. The required short text is composed of entire sentences selected from the documents of the collection which is called extractive summarization task. This task consists of weighting some sentences from the whole collection. We improve a language-independent method of weighting the sentences of the document collection looking for new terms which permits to improve results for multi-document task. Our results show that this method is among the best systems in the existing language-independent state-of-the-art methods.

1 Introduction

The text summarization tasks can be classified into single-document and multi-document summarization. In multi-document summarization, the summary of a whole collection of documents is built, in contrast to single-document summarization where the summary of only one document is to be built. So, a summary of a collection of documents is a single short text that contains the most important information from this collection of documents.

There are extractive and abstractive summaries. An abstractive summary is a short text that describes in different words the contexts of the source collection of documents [1]. Abstractive summarization process consists of re-phrasing the original collection of documents in fewer sentences. Usually, an abstractive summarization method uses linguistic methods to re-phrase and re-write documents. While this may seem better alternative to obtain a summary; really, in state-of-the-art, the abstractive methods offer worse quality than extractive methods.

An extractive summary is a selection of sentences or even paragraphs from the original collection of documents. It means, an extractive summarization method only decides, for each sentence, whether or not it will be included in the summary, and

* Work done under partial support of Mexican Government (CONACyT, SNI, PROMEP, SEP). The authors thank UAEM for its assistance.

then, it is usually presents to the user in the same order as in the source document collection. The resulting summary is read it too awkward; however, simplicity of the extractive summarization methods offers an attractive, robust and language-independent option, in contrast to more complicated abstractive methods.

A typical extractive summarization method consists in several steps [2], at each of them different options can be chosen. We will assume that the units of selection are sentences. Thus final goal of the extractive summarization process is sentence selection.

One of the ways to select the appropriate sentences is to assign some numerical measure of usefulness of a sentence for the summary and then select the best ones; the process of assigning these usefulness weights is called sentence weighting.

One of the ways to estimate the usefulness of a sentence is to sum up usefulness weights of individual terms of which the sentence consists; the process of estimating the individual terms is called term weighting. For this, one should decide what the terms are: for example, they can be words; deciding what objects will count as terms is the task of term selection. Different extractive summarization methods can be characterized by how they perform these steps.

For term selection step, several works employ domain-dependent terms like key-phrases, lexical chains, proper names and anaphors; and other ones employ word-based terms like words or n-grams [3-11]. Word-based terms have the advantage of depend less or nothing on the domain, since they are extracted from the source text to be summarized. Recently, the so-called Maximal Frequent Sequences (MFSs) have shown to be good terms for single-document extractive summarization with advantages over n-grams since each MFS only depends on the structure of the text, preserving more the sequential nature of the text [2, 12-14]. In each MFS not only is recovering the frequency of the term but also the size (number of words) of the term which provide more information about the importance of such term. Also, a maximal frequent sequence is a meaningful term since all its sequences are frequent too.

It is very useful the automatic construction of summaries for different collection of documents. For example, summaries of collection of news articles, sport reviews, research papers, description of products, politic debates, information from web pages, all today's news or all search results for a query. Such summaries present brief information about different points of view, opinions, sources of knowledge, controversial or agreement discussions on any event, fact, topic, history, book, web page, etc. It will help to the user not only quickly find the necessary information, but also understand what is the overall situation on a specific topic.

One possible scenario is when a researcher tries to understand a specific science area; in this case, the summary will return short definitions about this area. Another scenario, for example, when we want to know what books are in the library of a particular author. Given the name of the author, will be very interesting not only to see the names of the books of this author but also a short description of each book.

In this paper, we are experimenting with MFSs for multi-document task in order to improve the quality of extractive summaries.

The paper is organized as follows. Section 2 summarizes the state-of-the-art of multi-document summarization methods. In Section 3, some notions used for term selection in our method are introduced. Section 4 presents our experimental setting. Sections 5 and 6 describe the obtained experimental results for different term

selection and term weighting schemes, respectively, which are compared in Section 7 with those of existing methods. Section 8 concludes the paper.

2 Related Work

One of the most competitive methods of the-state-of-the-art methods are presented by Mihalcea [15] in the form of a clear graph-based formalism. In this method, the words that have closer relationships with a greater number of "important" words become more important themselves, the importance being determined in a recursive way similar to the PageRank algorithm used by Google to weight web pages. The latter idea can be applied directly to sentence weighting without term weighting: a sentence is important if it is related to many important sentences, where relatedness can be understood as, say, overlap of the lexical contents of the sentences [15]. The two methods presented in [15, 16] are those that currently give the best results and with which we compare our suggested method.

Another relevant works are developed by Wei *et al.* [17] where the relevance of a term is derived from an ontology constructed with formal concept analysis. Song *et al.* [3] weight a word calculating the number of lexical connections, such as semantic associations expressed in a thesaurus that the word has with its neighboring words; along with this, more frequent words are weighted higher. Passages are retrieved using a language model [18] with the objective to predict the probability of word sequences actually occur and low probability on word sequences that never occur. The n-gram model is used as a basis for the proposed language model.

A special procedure is designed by Nenkova *et al.* [19, 20] for comparative analysis of the content of several texts. In these works special terms are annotated using the pyramid scheme. The presence of each term in all documents of the collection accumulates the importance of this term. The more documents have the term, the more important is this term, and consequently will be included in the summary.

One of the most implemented sentence selection methods are supervised learning methods which consider sentence selection as a classification task. These methods train a classifier using a collection of documents supplied with existing summaries. As features of a sentence such methods can consider text units (in which case we can speak of term selection) or other, non-lexical characteristics. Different lexical and non-lexical features have been used in [21-23]. Most of these features are "heuristically motivated", since they tend to emulate the manual creation of extracts.

In a work of Kupiec [21], the following features were proposed: sentence position, sentence length, the presence of key phrases and overlap with the title of the document. More recent works [22, 23] extend these features incorporating information about the occurrence of proper names and the presence of anaphors. The "heuristically motivated" features allow extract very precise summaries. However, they have a very big disadvantage of being highly linked to a specific domain. This condition implies that the change for one domain to another, it may be necessary to redefine or even eliminate some features. For instance, key phrases, which are

particular for each domain, require being modified, while the overlap with the title, which has no sense in all topics, may be eliminated.

In order to increase the domain (and language) independence of machine learning summaries, Villatoro [24] eliminates all kind of "heuristically motivated" attributes and substitute them by word-based features. In particular, he uses word sequences (n-grams) as terms. Although the first attempt to use n-grams is exceeded the results of other methods, it has some disadvantages. One is that they are always sequences of a fixed size, which was previously defined by the user. The big part of the problem in such techniques lies in defining the size of the sequence to be extracted, which usually depends on the analysis of the text.

A very old and very simple sentence weighting heuristic in single-document summarization does not involve any terms at all: it assigns highest weight to the first sentences of the text. Texts of some genres—such as news reports or scientific papers—are specifically designed for this heuristic: *e.g.*, any scientific paper contains a ready summary at the beginning. This gives a baseline in single-document summarization [25] that proves to be very hard to beat on such texts. Similar to this heuristic in multi-document summarization is employed the first sentence of each document to conform the summary, each sentence is added until to reach the desired length [26]. However, comparing term-based methods with such position-based baseline is not fair in the sense that this baseline only works on text of specific genres and uses information (the position of the sentence) not available to term-based methods. It is worth noting that in Document Understanding Conference (DUC) competitions [25] only five systems performed above this baseline, which does not demerit the other systems because this baseline is genre-specific.

In a previous work [27], we analyzed several options for simple language-independent statistical term selection and corresponding term weighting, based on units larger than one word. In particular, we showed that so-called MFSs, as well as single words that are part of bigrams repeated more than once in the text, are good terms to describe a collection of documents. In this paper, we experiment with some minimum-frequency thresholds of MFSs in order to improve the quality of the multi-document summaries.

3 Proposed Method

The research on multi-document summarization is less developed than single-document summarization because summarizing a collection of thematically related documents is more difficult, than summarizing a single text. In order to avoid repetitions, one has to identify and locate thematic overlaps. In other words, has to locate the most important terms for representing a collection of documents. One also has to decide what sentences are more important, and to arrange events from various sources along a single timeline.

One of our hypotheses is that MFSs with higher threshold should generate summaries with better quality than MFSs with lower threshold. It can be explained on reason that there would exist in the language a multiword expression that can express

the same content in the more compact way which can be detected more precisely using higher threshold (see Experiment 1).

Another hypothesis is, in contrast to MFSs, FSs is important if there would exist in the language a single word or at least an abbreviation to express it. Such single words or abbreviations should be considered as bearing the more important meaning with lower threshold because we need to extract more single words or abbreviations to know if they can be used for composing a summary (see Experiment 2)..

The third hypotheses we explore in this work, is that MFSs represent in a better way the summarized content of collection of documents than FSs because their (MFSs) probability to bear important meaning is higher. It can happen because there are too many non-maximal FSs in comparison to MFSs (see Experiment 3).

Our proposed method followed the sequence of steps as follows:

1. Term selection: the main idea is to utilize the MFSs and its derived FS as the main terms.

1.1 Frequency threshold

2. Term weighting: each MFS can be weight based in its frequency or in its length;
3. Sentence weighting: we test the option calculating the sum of the weights of the terms contained in the sentence.
4. Sentence selection: we test the option when the sentences with greater weight are selected until the desired size of the summary (100 words) is reached.

3.1 Term Selection

An n-gram is a sequence of n words. We say that an n-gram occurs in a text if these words appear in the text in the same order immediately one after another. We call an n-gram frequent, if it occurs more than β times in the text, where β is a predefined threshold. Frequent n-grams—we will also call them frequent sequences (FSs)—often bear important semantic meaning: they can be multiword expressions (named entities: *The United States of America*, idioms: *kick the basket*) or otherwise refer to some idea important for the text (*the President's speech*, *to protest against the war*).

An n-gram can be a part of another, longer n-gram. All n-grams contained in an FS are also FSs, for example, the if MFS is *The United States of America* then *The United States* or *States of America* is a FS too. In this case, *The United States* tends to be synonymous to the longer expression, and the author of one document would choose one or another way to refer to the same entity. FSs that are not parts of any other FS are called Maximal Frequent Sequences (MFSs) [28, 29]. For example, in the following collection of documents

D1: ... *Mona Lisa* is the most beautiful picture of Leonardo da Vinci ...

D2: ... *Eiffel tower* is the most beautiful tower ...

D3: ... *St. Petersburg* is the most beautiful city of Russia ...

D4: ... *The most beautiful* church is not located in Europe ...

the only MFS with $\beta = 3$ is *is the most beautiful*, while the only MFS $\beta = 4$ is *the most beautiful* (it is not an MFS with $\beta = 3$ since it is not maximal with this β). As this

example shows, the sets of MFSs with different thresholds do not have to, say, contain one another.

The notions of FSs and MFSs are closely related to that of repeating bigrams. This set is conceptually simpler, but for computational implementation MFSs could be more compact.

For term selection, we compared MFSs with more traditional word-based features such as single words and n-grams. Namely, we considered the following variants of term selection:

- M : the set of all MFSs with some threshold β . In the example from Section 3, $M = \{\text{is the most beautiful}\}$. Also, we denote by M_2 the set of all MFSs with $\beta = 2$.
- W : single words (unigrams) from elements of M . In our example, $W = \{\text{is, the, most, beautiful}\}$.

Optionally, stop-words were eliminated at the pre-processing stage; in this case our bigrams (or MFSs) could span more words in the original text.

3.2 Term Weighting, Sentence Weighting and Sentence Selection

For term weighting, the frequency of the term was used; for sentence weighting, the sum of the weights of the terms contained in the sentence was used; for sentence selection, the sentences with greater weight were selected until the desired size of the summary (100 words) is reached.

Optionally, stop-words were eliminated at the pre-processing stage. For term weighting, different formulae were considered containing the following values:

- f : frequency of the term in MFSs, i.e., the number of times the term occurs in the text within some MFS. In our example, $f(\text{is}) = 3$ since it occurs 3 times in the text within the MFS *is the most beautiful*. Under certain realistic conditions (MFSs do not intersect in the text, words do not repeat within one MFS) f is the number of times the term occurs in the text as part of a repeating bigram. In our example, $f(\text{is}) = 3$ since it occurs 3 times in a repeating bigram *is the* (and one time in a non-repeating context *church is not*).
- l : the maximum length of an MFS containing the term. In our example, $l(\text{is}) = 4$ since it is contained in a 4-word MFS *is the most beautiful*.
- 1 : the same weight for all terms.

For sentence weighting, the sum of the weights of the terms contained in the sentence was used. For sentence selection, the following options were considered:

- best: sentences with greater weight were selected until the desired size of the summary (100 words) is reached. This is the most standard method.

4 Experimental Setting

We realized several experiments in order to verify our hypotheses formulated in the previous section. The specific settings for each step varied between the experiments and are explained below for each experiment.

Test data set. We used the DUC collection provided [25]. In particular, we used the data set of 60 document collections which consist of 567 news articles of different length and with different topics. Each collection of documents in the DUC collection is supplied with a set of human-generated summaries provided by two different experts. While each expert was asked to generate summaries of different length, we used only the 100-word variants.

Evaluation procedure. We used the ROUGE evaluation toolkit [30] which was found to highly correlate with human judgments [31]. It compares the summaries generated by the program with the human-generated (gold standard) summaries. For comparison, it uses n -gram statistics. Our evaluation was done using n -gram (1, 1) setting of ROUGE, which was found to have the highest correlation with human judgments, namely, at a confidence level of 95%.

As a kind of statistical significance check, we randomly divided our test data into two halves and ran this (and most of the other) experiments separately on each subset. These experiments confirmed the qualitative observations reported in this paper.

Previous results. We tried term selection options, such as M and W , with the term weighting option l , l , the options related to f , and their combination (Table 1). For sentence selection, we tried the *best* combination. Term selection W gave a slightly better result than M . The best results are highlighted in boldface. (See more details in [27]. We borrow this table from [27] to compare with Tables 2-4).

Table 1. Results for different term selection and term weighting options for multi-document summarization.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
M	F	best	0.31372	0.31986	0.31660
	f^2		0.31162	0.31870	0.31499
	l		0.30620	0.31347	0.30965
	L		0.31411	0.32199	0.31786
	l^2		0.29184	0.30275	0.29706
	$f \times l$		0.31329	0.32103	0.31696
	$f \times \times l$		0.28328	0.29592	0.28933
W	F	best	0.31919	0.32494	0.32192
	l		0.26413	0.27828	0.27072
	f^2		0.30056	0.30764	0.30391

Experiment 1. For this experiment, we use the configuration of the algorithm of previous results (see Table 1) of this section. Then we tested the algorithm with $\beta = 2$ (see Table 2).

Table 2. Results for different term selection and term weighting options with $\beta = 2$.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
<i>M</i>	<i>F</i>	best	0.29038	0.29749	0.29373
	<i>f²</i>		0.29038	0.29749	0.29373
	<i>I</i>		0.29038	0.29749	0.29373
	<i>L</i>		0.29881	0.30714	0.30279
	<i>I²</i>		0.28837	0.29912	0.29351
	<i>f × I</i>		0.29881	0.30714	0.30279
	<i>f × × I</i>		0.28455	0.29606	0.29006
<i>W</i>	<i>F</i>	best	0.32238	0.32902	0.32557
	<i>I</i>		0.26413	0.27828	0.27072
	<i>f²</i>		0.29559	0.30361	0.29796

Experiment 2. For this experiment, we use the configuration of the algorithm of previous results (see Table 1) of this section. Then we tested the algorithm with $\beta = 3$ (see Table 3).

Table 3. Results for different term selection and term weighting options with $\beta = 3$.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
<i>M</i>	<i>F</i>	best	0.30601	0.31302	0.30933
	<i>f²</i>		0.30601	0.31302	0.30933
	<i>I</i>		0.30601	0.31302	0.30933
	<i>L</i>		0.31574	0.32286	0.31911
	<i>I²</i>		0.29904	0.30852	0.30356
	<i>f × I</i>		0.31574	0.32286	0.31911
	<i>f × × I</i>		0.28309	0.29270	0.28768
<i>W</i>	<i>F</i>	best	0.32279	0.32826	0.32538
	<i>I</i>		0.26413	0.27828	0.27072
	<i>f²</i>		0.30934	0.31602	0.31253

Experiment 3. For this experiment, we use the configuration of the algorithm of previous results (see Table 1) of this section. Then we tested the algorithm with $\beta = 4$ (see Table 4). Comparison of the results for the proposed method is shown in Tables 5 and 6. The results of the state-of-the-art methods [25] for multi summarization are shown in Table 7.

Table 4. Results for different term selection and term weighting options with $\beta = 4$.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
<i>M</i>	<i>F</i>	best	0.30964	0.31648	0.31288
	<i>f</i> ²		0.30964	0.31648	0.31288
	<i>I</i>		0.30964	0.31648	0.31288
	<i>L</i>		0.32326	0.32980	0.32636
	<i>f</i> ²		0.30825	0.31812	0.31296
	<i>f</i> × <i>I</i>		0.32326	0.32980	0.32636
	<i>f</i> × × <i>I</i>		0.29651	0.30533	0.30079
<i>W</i>	<i>F</i>	best	0.31855	0.32329	0.32076
	<i>I</i>		0.26413	0.27828	0.27072
	<i>f</i> ²		0.30934	0.31602	0.31253

5 Conclusions

We observed that MFSs with higher threshold generate summaries with better quality than MFSs with lower threshold. It can be explained on reason that there exist in the language multiword expressions that can express the same content in the more compact way which can be detected more precisely using higher (see Table 5).

Then, we observed that, in contrast to MFSs, FSs is important if are extracted with lower threshold. It can be explained because there exist in the language a lot of single word or at least an abbreviation to express an important meaning.

Such single words or abbreviations should be considered as bearing the more important meaning with lower threshold because we need to extract more single words or abbreviations for knowing if they can be used for composing a summary (see Experiment 2).

The third hypotheses we explore in this work, is that MFSs represent in a better way the summarized content of collection of documents than FSs because their (MFSs) probability to bear important meaning is higher. It can happen because there are too many non-maximal FSs in comparison to MFSs (see Experiment 3).

Table 5. Comparison of results using different thresholds (terms are MFS).

Method	Recall	Precision	F-measure
M where $\beta = 2, 3, 4$	0.31411	0.32199	0.31786
M where $\beta = 2$	0.29881	0.30714	0.30279
M where $\beta = 3$	0.31574	0.32286	0.31911
M where $\beta = 4$	0.32326	0.32980	0.32636

Table 6. Comparison of results using different thresholds (terms derived from MFS).

Method	Recall	Precision	F-measure
W where $\beta = 2, 3, 4$	0.31919	0.32494	0.32192
W where $\beta = 2$	0.32238	0.32902	0.32557
W where $\beta = 3$	0.32279	0.32826	0.32538
W where $\beta = 4$	0.31855	0.32329	0.32076

We compared the following results (see Table 7):

- **State of the art:** The best top 5 systems from 17 systems in DUC 2002 for multi-document summarization task are listed in Table 7 [25].
- **Baseline:** DUC collection has a configuration denotes as *Baseline*, which selects the first sentence in the first, second, third, and so on document in chronological sequence until you have the target summary size [25]. This baseline gives good results on the kind of texts (news reports) that we experimented with. Thus we can compare with this configuration of baseline. Also we believe this configuration to be a more realistic baseline for the types of texts.
- **Recent work:** As shown in Table 1 (see [27]), using the configuration the W term selection scheme and the f term weighting scheme. We call it as the 4th best method.
- **Our proposal:** We compare these methods with the best results obtained in this paper: M term selection scheme with $\beta = 4$ and the l term weighting scheme, as shown in Table 5. W term selection scheme with $\beta = 2$ and the f' term weighting scheme, as shown in Table 6.

We tested new method for the automatic generation of text summaries for a multi-document summarization based on the discovery of MFSs, specifically we tested different combinations of term selection, term weighting, sentence weighting and sentence selection schemes with different thresholds. Comparing to other methods, we did not receive the best results but considering that the proposed method is language- and domain-independent, we think that the results are very encouraging. Also we improve the obtained results using MFSs (see 4th best method and proposed in Table 7).

Table 7. Comparison of results with other methods.

Method	F-measure
1st best method	0.3578
2nd best method	0.3447
Best proposed	0.3264
3rd best method	0.3264
4th best method	0.3219
5th best method	0.3056
6th best method	0.3047
Baseline	0.2932

References

1. Lin, C.Y. and Hovy, E. Automated Text Summarization in SUMMARIST. In *Proc. of ACL Workshop on Intelligent, Scalable Text Summarization*, Madrid, Spain, 1997.
2. Ledeneva Y., Gelbukh A., García-Hernández R. Terms Derived from Frequent Sequences for Extractive Text Summarization. LNCS 4919, pp. 593-604. Israel, Springer-Verlag, ISSN 0302-9743, 2008.
3. Song, Y., et al. A Term Weighting Method based on Lexical Chain for Automatic Summarization. *CICLing 2004*, LNCS, vol. 3878, Springer-Verlag 2004.
4. Cristea D., et al. Summarization through Discourse Structure. *CICLing 2005*, LNCS, vol. 3878, Springer-Verlag 2005.
5. Liu, D., He, Y., Ji, D., Hua, J. Multi-Document Summarization Based on BE-Vector Clustering. *CICLing 2006*, LNCS, vol. 3878, Springer-Verlag 2006.
6. Xu, W., Li, W., et al. Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis. *CICLing 2006*, LNCS, vol. 3878, Springer-Verlag 2006.
7. HaCohen-Kerner, Y., Zuriel, G., Asaf, M. Automatic Extraction and Learning of Keyphrases from Scientific Articles. LNCS, vol. 3878, pp. 645-657, Springer-Verlag 2005.
8. Gelbukh A., Sidorov G., Sang Yong Han, Hernandez-Rubio E. Automatic Enrichment of Very Large Dictionary of Word Combinations on the Basis of Dependency Formalism. In: Raúl Monroy, et al. (Eds.). ISSN 0302-9743, vol. 2972, pp. 430-437, 2004.
9. Sidorov, G., Gelbukh A., *Procesamiento Automático del Español con enfoque en recursos léxicos grandes*. IPN, ISBN 970-36-0264-9, 2006.
10. Bolshakov, I.A. Getting One's First Million... Collocations. *CICLing-2004*, Seoul, Korea, LNCS 2945, p. 229-242, Springer-Verlag 2004.
11. Filippova K., Mieskes M., Nastase V., et al. Cascaded Filtering for Topic-Driven Multi-Document Summarization. *Proc. of Document Understanding Conference 2007*. <http://duc.nist.gov/pubs.html#2007>.
12. Ledeneva Y. Effect of Preprocessing on Extractive Summarization with Maximal Frequent Sequences. LNAI 5317, pp. 123-132, Springer-Verlag, ISSN 0302-9743, 2008.
13. García-Hernández R., Ledeneva Y., Gelbukh A., Rendon E., Cruz R. Text Summarization by Sentence Extraction Using Unsupervised methods. LNAI 5317, pp. 133-143, Mexico, Springer-Verlag, ISSN 0302-9743, 2008.
14. Ledeneva Y., Gelbukh A., García-Hernández R. Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. In: G. Sidorov et al (Eds). *Advances in Computer Science and Engineering, Research in Computing Science*, vol. 34, pp. 163-174, ISSN 1870-4069, 2008.
15. Mihalcea, R. Random Walks on Text Structures. *CICLing 2006*, LNCS, vol. 3878, pp. 249-262, Springer-Verlag 2006.
16. Mihalcea, R., Tarau, P. TextRank: Bringing Order into Texts, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, 2004.
17. Xu Wei, Li Wenjie, et al. Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis. *CICLing 2006*, LNCS, vol. 3878, Springer-Verlag, pp. 480-489, 2006.
18. Ying J.-C., Yen S.-J., Lee Y.-S. Language Model Passage Retrieval for Question-Oriented Multi Document Summarization. *Proc. of Document Understanding Conference 2007*. <http://duc.nist.gov/pubs.html#2007>.
19. Nenkova A. Understanding the process of multi-document summarization: content selection, rewriting and evaluation. Ph.D. Thesis, Columbia University, 2006.
20. Nenkova A., Passonneau R. Evaluating content selection in summarization: The pyramid method. In: *Proc. of NLT/NAACL-2004*, 2004.

21. Kupiec, J., Pedersen, J.O., Chen, F. A Trainable Document Summarizer. In *Proceedings of the 18th ACM-SIGIR Conference on Research and Development in Information Retrieval*. Seattle, pp. 68-73, 1995.
22. Chuang T.W., Yang J. Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms. *Proc. of the ACL-04 Workshop*. Barcelona, España, 2004.
23. Neto L., Freitas A., et al. Automatic Text Summarization using a Machine Learning Approach. *Proc. of the ACL Workshop*, España, 2004.
24. Villatoro-Tello, E., Villaseñor-Pineda, L., Montes-y-Gómez, M. Using Word Sequences for Text Summarization. *TSD, LNAI Springer* 2006.
25. DUC. Document understanding conference 2002; www-nlpir.nist.gov/projects/duc.
26. Lin Chin-Yew and Eduard Hovy, From Single to Multi-document Summatization: A Prototype System and its Evaluation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic (ACL)*, 2002. pp.457-464.
27. Ledeneva, Y., García-Hernández, R.A., Gelbukh A. Multi-document Summarization using Maximal Frequent Sequences. *Research in Computer Science*, pp.15-24, vol. 47, ISSN 1870-4069, 2010.
28. García-Hernández, R.A., Martínez-Trinidad J. F., Carrasco-Ochoa J. A. A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text, *CIARP'2004, LNCS vol. 3287 Springer-Verlag* 2004. pp. 478-486.
29. García-Hernández, R.A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A. A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection, *CICLing'2006, LNCS vol. 3878 Springer-Verlag* 2006. pp. 514- 523.
30. Lin C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004.
31. Lin C.Y., Hovy E. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of HLT-NAACL, Canada*, 2003.

Establishing a Software-Subcontracting Management Model to Improve the Software-Subcontracting Process in Small-size Enterprises

Garcia, I. and Pacheco, C.

Postgraduate Department
Technological University of the Mixtec Region, Mexico (www.utm.mx)
{ivan@mixteco.utm.mx, dago@mixteco.utm.mx}

Abstract. Software subcontracting using external suppliers could be the best option in reducing the investment and implementation costs of new solutions, and would allow an organization to use their resources more efficiently. Two centuries ago John Ruskin said: "the quality never is an accident; always is the result of an intelligence effort" Subcontracting is an intelligent decision, and particularly now when small-size enterprises are facing the major economic power of large corporations. The subcontracting process offers a 'nearby colleague' to increase the benefits and quality of service of the enterprises, at the same time reducing the investment. This paper illustrates the design of an alternative model to improve the software-subcontracting management process. The model is orientated towards to small-size enterprises, as the high implementation costs of the methods reviewed in this study are well known. As additional work, this paper depicts the current situation of software-subcontracting processes, reviewing the principal models applied in industry, and establishes a base comparative of the developed model. The results obtained by the model's implementation are presented to expose its applicability.

Keywords: Software-subcontracting process, service model, effective practices, small-size enterprises.

1 Introduction

The software-subcontracting process is defined as the process of acquiring, either partially or totally, Information Technology (IT) from an external services supplier [17]. It means delegating, via a contract, all or part of the software work to an external company that joins in the client organizational strategy and seeks to design a solution to existing problems within this. In recent years, software subcontracting has been of increasing interest to researchers and a considerable number of organizations. In spite of its origins in the 60's and 70's [24], the subcontracting process was only applied to financial areas and support operations; nowadays the organizations are focused on saving costs; paying more attention to their essential business processes; accomplishing economic, strategic and/or technological advances; or, basically, determining if their IT functions are unnecessary, ineffective or incompetent. It is important to say that not until the Eastman Kodak study [25] was presented in 1989, was the subcon-

tracting process considered as an efficient strategy for organizations to select, and emerged as one of the ten fundamental issues for their survival. However, instead of the successful experiences that had been published [5] [25] [14] [4] [28], the projects subcontracting failures still remain within organizations because the organizations that subcontract loose control of the acquired software products/services, as well as the usual problems of delayed delivery, costs exceeded, the poor quality of subcontracted products, and more [13].

The remainder of the paper is organized as follows. Section 2 presents the background to software-subcontracting process that established the motivations for proposing our alternative model. The principal models and standards used to manage the software-subcontracting process are analyzed in Section 3. Section 4 presents an empirical comparative between models to emphasize the significant disadvantages for applying them in small enterprises. Section 5 presents a detailed description of the proposed model. To demonstrate the feasibility of our model, the experimental results are explained in Section 6. Finally, conclusions and future work are drawn in Section 7.

2 Motivations

Some small companies have had good experiences subcontracting for example the payroll and accounting management departments, leaving these services in professional hands and obtaining important cost reductions. But, if these companies have good results subcontracting those processes, why do they not also delegate software services and processes to an external supplier? Nowadays, the principal disadvantages to subcontracting software services are fundamentally the lack of guidance for its implementation, the need to commit to dedicating important long term resources, and the Return of Investment (ROI) which is just too much to expect for many companies [12] [20]. The fast rate of technological innovation in IT impedes many companies from being up-to-date, but they do not want to 'miss the technological train' and the definition of alternative methods is necessary to avoid the system's obsolescence [8]. Nevertheless, the software-subcontracting process is complex because it is managed externally in order to acquire products, systems and services, rather than internally, where they could manage their own processes. As a result, the identification of effective practices in the delegation of responsibilities of the software services and processes are focused internally, to ensure that the subcontracting process will be effective, and externally, in order that the company can manage the subcontracting process and take control of their suppliers. These effective practices provide the fundamental component of a subcontracting discipline and the rigor that enables the rapid development of products/services with a high level of success [3].

This approach requires a renewed dedication to ensure that a defined, implemented, measured and maintained software-subcontracting process exists. The implementation of an effective software-subcontracting process not only makes the technological difference in an organization with regard to its competitors, it also facilitates the concentration of their own resources in the 'core business', focusing on the activity for which the organization was created, generating higher benefits without

losing the capacity of maintaining service quality for clients and identifying new business opportunities [9].

3 Related Work

This Section aims to show a brief review of the most significant commercial models and standards used to manage the software-subcontracting process in big software organizations. Unfortunately, most models could not be implemented in small enterprises because they were created for big organizations. Besides, small-size enterprises usually confuse two terms; "acquiring" and "subcontracting" and try to adopt a model without a clear understanding of both terms. Currently, there are two approaches to guide small enterprises in adopting an efficient software-subcontracting process: to manage and monitor the subcontracting process effectively according to a defined standard; to establish deliverable products and staff roles and responsibilities, etc.; and to adapt commercial models within an improvement initiative. Almost all small enterprises fail when implementing these models, but they still use them; so, it is convenient to analyze each one and try to identify a set of effective practices.

One of the process areas included in CMMI-DEV 1.2 is the Supplier Agreement Management (SAM) process area [6]. The purpose of Supplier Agreement Management is to manage the acquisition of products from suppliers through a formal process. According to its description, the model provides seven practices to perform the SAM process. This 'managed' process begins with the identification of the acquisition type to perform and closes with the transition of the subcontracted products to the project. The research by [26] presents a business workflow process model for SAM process area of CMMI: Capability Level 2. It consists of three layers: contextual layer, elaboration layer, and definition layer. A software tool called Supplier Agreement Management Tool was also developed to help integrate the details of this approach. However, the fundamental disadvantage of CMMI models is the high cost of implementation, unaffordable for small enterprises. Small enterprises must obtain a "certification" of CMMI Capability Levels for each process that they do not consider a strategic decision; so, it is very difficult to address the software subcontracting problems with this excessively formal model.

The ISO/IEC 15504:2004 standard goes beyond quality audits to help an organization assess how well its processes perform. One standard's objective is to guide organizations in software-subcontracting process through the determination of the potential supplier's capacity [15]. This evaluation enables risk identification, related to each supplier, when an organization subcontracts software products or services. The purpose of ISO/IEC 15504:2004 subcontracting and supply processes is to obtain the product and/or service that satisfies the need expressed by the customer (the standard assumes this is a software oriented product or service - this is apparent in the supply process). According to the standard, the software-subcontracting process begins with the identification of a customer requirement and ends with the acceptance of the product and/or service. However, ISO/IEC 15504:2004 is more known as a Software Process Improvement model, but there is no strong evidence that it had been useful in implementing software-subcontracting processes in small enterprises. Besides,

process dimension should be wider and covers all possible lifecycles applicable in small enterprises; it is impossible that all process attributes be universal and can be used by all processes and base practices without an expensive cost.

In SA-CMM an individual subcontracting process begins with the definition of a customer need and ends with the contract closure. SA-CMM is designed to be sufficiently generic for use by any government or industry organization, regardless of size, for subcontracting products. According to the SEI [21], effective subcontracting processes are critical to the success of process improvement, but the output quality is only determined within the context of organizational business-needs. There is relevant research using SA-CMM in industrial environments. Wong [29], for example, summarized software-subcontracting management lessons learned from a complex multidisciplinary and contract environment. He also identified a number of measures for improvement in a project. Later, these measures were analyzed against SA-CMM Process Areas for their applicability and comprehensiveness. However, the "problem" for demonstrating the small enterprises' applicability endures.

COBIT provides good practices for the management of IT processes in a manageable and logical structure, meeting the multiple needs of enterprise management by bridging the gaps between business risks, technical issues, control needs and performance measurement requirements. Information systems subcontracting, development and maintenance should be considered in the context of the organization's IT long and short-term plans. The organization's system development life cycle methodology should provide for a software-subcontracting strategy plan, by defining whether the software will be acquired off-the-shelf, developed internally, through contract or by enhancing existing software - or a combination of all of the above [7]. But more than a guideline of implementation, COBIT is designed to help in understanding and managing the risks and benefits associated to information and IT related. It is not a model that focuses exclusively in the software-subcontracting process and it has a noticeable tendency towards the supervision of the IT generic processes.

ITIL's Availability Management provides reliable access to IT services. Availability means that the client will always receive the expected services as necessary. The main benefits are: supplier performance improvement and detailed information availability for negotiations at service level [19]. In spite of the use of ITIL in small enterprises, there are only three processes that are not completely covered by the model: Supplier Management (related to software-subcontracting process), Business Relation Management, and Service Report. The small enterprise should be very carefully if tries to implement some process from ITIL, because it requires a lot of time and money without warranty of success.

The IEEE Recommended Practice for Software Acquisition 1062 recommends a set of useful quality practices that can be selected and applied during one or more steps in a software acquisition process or software-subcontracting process. According to standard's documentation, IEEE 1062 is designed to help organizations and individuals to incorporate quality considerations during the definition, evaluation, selection, and acceptance of supplier software for operational use; and to determine how supplier software should be evaluated, tested, and accepted for delivery to end users [22]. However, the most common problem with the IEEE 1062 standard is the lack of mechanism in planning the subcontracting-project and to elicit and monitor the subcontracting requirements. Using the standard is too difficult for small enterprises be-

cause it requires a lot of documentation and assumes that the organization knows the standard very well.

As we see, there is wide research in developing models and standards to establish efficient software-subcontracting process. However, almost all models and standards are not affordable for small-size enterprises but provide a set of effective practices that could be customized in a simple model.

4 An Empirical Comparative of Subcontracting Models

Some research and critique has already been done into the reviewed models and standards; important criteria to be considered when developing an alternative approach. Research by [16], for example, discussed the implementation of CMMI in small enterprises and showed the following disadvantages: there is not a customized guideline; continuous representation enables the selection of only those process areas where the organization feels comfortable; an exponential increase in areas and practices, time, resources and cost; an excessive standardization for small organizations that work and evolve in different ways to large ones.

Research by [27] provides some important information about the standard ISO/IEC 15504:2004. Reaching the capacity dimension is a big difficulty in small enterprises; there is overlap with the process dimension. Assessment complexity (and its cost) is significantly higher than other models. In the context of the software-subcontracting process, the ISO standard does not provide a formal guide to monitoring a product when it has been accepted.

SA-CMM has a high level of complexity; the period of adoption in small enterprises is longer than in larger enterprises. In this case, the statement of "*simplifying activities does not affect the professional level*" does not apply [2] [29].

More than an implementation guideline, COBIT is designed to help organizations to understand and manage the risks and benefits associated with information and related IT. COBIT is not a model that exclusively focuses on the software-subcontracting process and shows a clear tendency to supervise the IT processes.

With respect to ITIL, there are three processes that are not explicitly covered: Subcontracting Management, Business Relation Management, and Service Level Report. These processes are ignored and simultaneously fitted within the ITIL frame, for example the Subcontracting Management is included in the Service Level Management. On the other hand, the experts notice that ITIL has a list of minimum requirements against which an organization can be evaluated but it does not indicate how to reach a level of "conformity".

Despite their efficiency, the reviewed models make the same errors:

- The models specify what activities perform, however they do not provide guidelines about how to do it, in both engineering and project management.
- The models do not provide procedures for internal project management which should include own templates to facilitate the process, and fundamentally.
- There is no a customized guideline for small enterprises.

To summarize, in Table 1 we provide a homogenized analysis of the models; identifying important characteristics that we consider relevant in proposing an alternative model.

Some authors, for example [23], consider that standards reduce the developer's autonomy in large enterprises; developers are overloaded with extra work, assume coercive restrictions that suffocate the creativity required in innovating software development, and focus purely on the process, ignoring the people involved.

However, there is one that defends interdependence (as opposed to independent work) and takes a collaborative form [1], mature levels of process make a development process more 'socialized' where the collaborative effort increases efficiency and effectiveness (in our case when companies try to subcontract software as business strategy). We try to explore the benefits of the last approach, adapting it in small enterprises.

Table 1. Comparative analysis over reviewed models and standards

Criteria	Models and Standards					
	CMMI-DEV	ISO/IEC 15504	SA-CMM	ITIL	COBIT	IEEE 1062
Use of procedures for internal supplier management	X	✓	✓	X	X	✓
Use of templates for service management	X	X	X	X	X	✓
Establish a Service Level Agreement	X	X	X	✓	X	X
Incorporate an on going process improvement	✓	✓	✓	✓	X	X
Use of metrics for management	✓	✓	✓	X	✓	X
A contract requirement from start to finish	✓	✓	✓	✓	✓	✓
Relevance of the customer/supplier relationship	✓	✓	✓	✓	X	✓
Customized for small enterprises	X	X	X	X	X	X
High cost of implementation	✓	✓	✓	✓	✓	✓

5 Definition of a Customized Model for Small Enterprises

None of the previous models were created to address the particular small enterprises necessities. We propose an alternative model for managing the software applications/services that any company subcontracts, the Software-Subcontracting Management Model (SSMM). This model attempts to cover the strengths and weaknesses determined in Table 1.

This model enables the organization to improve its software-subcontracting management process. We start with the idea that to perform service management, it is necessary to define the activities/tasks to execute; when these should be performed; and what entries and outputs should be obtained. Our model is a guideline for suppliers and helps them in the project transference and operative service procedures execution. SSMM stages insist on those aspects that have major difficulties. The service management that our model offers is based on the efficient coordination of the three "Ps": *people*, *processes* and *products* through the three stages depicted in Table 2.

Table 2. Phases and generic objectives of SSMM

Stage name	Objectives per Stage
Initial stage	This should be planned with the client. A Project Plan must be established.
Stabilization stage	Supplier takes total responsibility for service according to Service Level Agreements (SLA). The duration is set in the contract and is "specific" for each client.
Closure stage	The closure of the service is prepared and the transition to the client is performed. A formation plan should be developed to prepare those personnel which assume the service maintenance and continuity.

SSMM helps to define the Software Subcontracting Management process tasks, identifying the correct moment for its application and recognizing the inputs and outputs required by the process. This model begins when the supplier is in charge of the service and finishes when the contract finalizes. SSMM is a guide for the operative transference of the project and it focuses on those aspects detected as deficient and incomplete, reducing the stages of the life cycle [5], [17], [18], [10] from four to three. Previous research determined that many subcontracting practices had not been used by the assessed small enterprises. Now, we are trying to define and implement effective practices according to the real characteristics and necessities of these very important organizations.

Each stage of SSMM defines a set of effective practices that it is necessary to accomplish to continue to the next stage. The model also provides a set of activities examples to help the companies to implement the methodology, and includes an expected list of work products for each stage (see Figure 1). The model stages are defined as follows:

- **Initial Stage.** In this stage the SLA is established, and the timescale for this is not less than one month and no more than three months. Of course this depends on the scope and the complexity of the service to be subcontracted. The activities of the stage are planned with the client, producing a Project Plan and the SLA. The sequence of activities is neither "rigid" nor "sequential", as it is possible to perform these in parallel.
- **Stabilization Stage.** The external supplier takes control of the services previously defined in the SLA. The stage duration is established in the contract (in years) and is specific to each client.
- **Closing Stage.** The objective of the stage is to prepare the conclusion of the service and perform the transference of the service to the client or a third party supplier. This stage covers the following aims:
 - To perform the activities of evolutionary and corrective maintenance and/or new developments, in such a way that the SLA is not affected.
 - To plan the service transference defining the assumptions and conditions, and economic impact of the same.
 - To analyze the services to return from the point of view of the complexity of resources, significance and availability.
 - To develop a training plan oriented to the personnel that will assume the continuity of maintenance.
 - To transfer existing knowledge in documents, records, and more.
 - To transfer sources, libraries, supports and services.

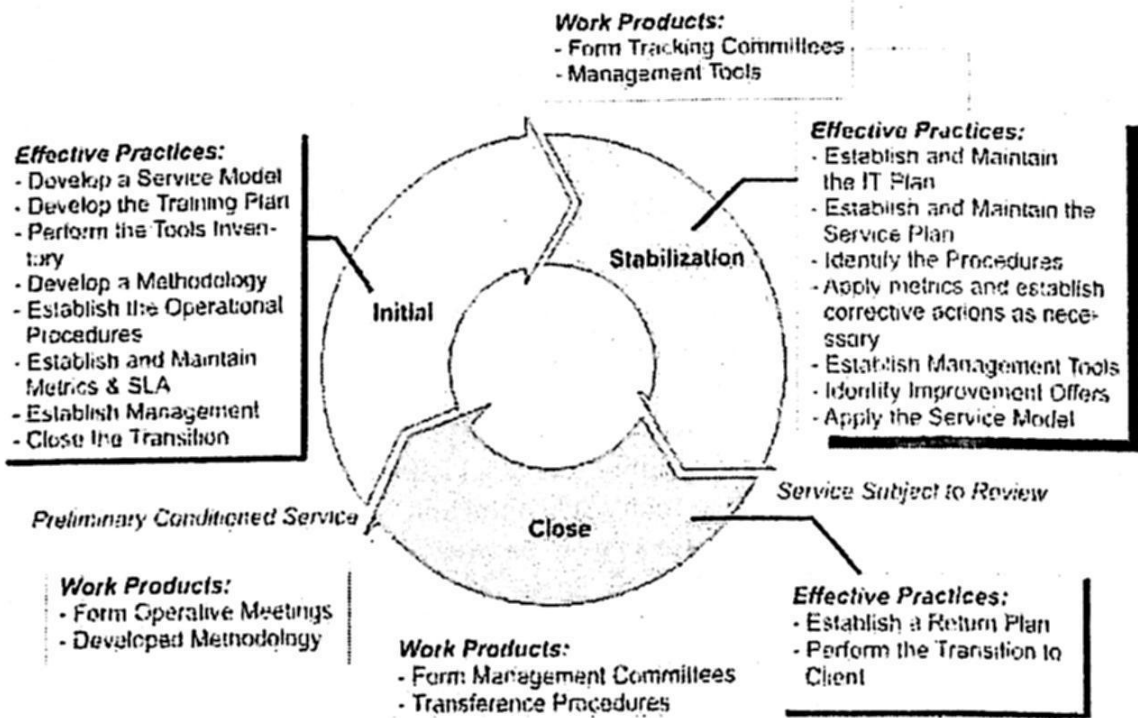


Fig. 1. SSMM stages

5.1 The SSMM structure.

We believe that adequate management of the subcontracting process needs a set of elements that it is necessary to implement during the Transition stage, incorporated in SSMM. These elements can be provided by both the client and the supplier. However it is recommended that it comes from the providing company in order to obtain homogeneous services. Figure 2 shows the general SSMM structure.

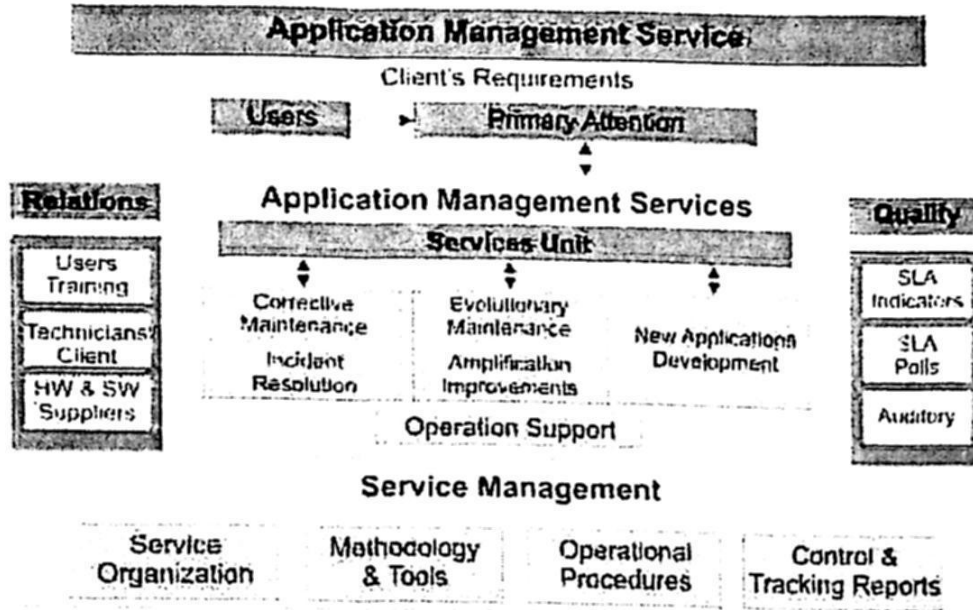


Fig. 2. SSMM structure.

We provide a brief description of SSMM elements:

- **Primary Attention.** This is the communication channel for managing incidents and requests between final users and the Services Unit. Commonly known as the Help Desk.
- **Services Unit.** This is the set of technical personnel in charge of service delivery. Among the developed activities are:
 - **Corrective Maintenance.** Activities for incidents and failures resolution produced in the software under operation.
 - **Evolutionary Maintenance.** Activities for the accomplishment of modifications and improvements in the software under operation.
 - **New developments.** Activities for management and implementation of new software projects.
 - **Operational Support.** Activities for the resolution of specific or information management requests not structured as available functionality for the final user.
- **Relations.** These relations determine the responsibility matrix in the different activities carried out during the service: Users Training, Client Technical Personnel and HW and SW Suppliers.

- **Quality.** The quality enables performing the objective or subjective evaluation of the performed services.
- **Service Organization.** This describes the terms and conditions of the agreement between client and supplier, by which the latter company will provide IT services during the Ongoing Stage. The services scope and company responsibilities would be specified.
- **Methodology.** This systematically defines the way “to do things” in the software life cycle.
 - **Support Tools.** The set of client tools and supplier management tools, or a third party that can complement them.
- **Operational Procedures.** These define the operative flows between different client areas and the service unit.
- **Control and Tracking Reports.** These are the sets of generated reports that realize the service control and tracking.

6 Testing the Model

To evaluate the applicability of our model, we have implemented an improvement initiative within one Mexican small enterprise. TICDES® Software is a privately owned small software development enterprise with about twenty software engineers, five administrative staff, and two project leaders. This organization specializes in software development and software consultancy. The size of the projects that TICDES® has developed in the past are up to 35,150 LOC with a duration ranging from a couple of weeks to 4,000 hours. The number of people participating in the teams has ranged from three to seven. From the beginning, the company was created with the goal and commitment to develop quality software. Before this research, the company had not decided to launch an improvement initiative because of the lack of information on it. The current software demand exceeds the enterprise capabilities and a software subcontracting process is an alternative solution to avoid the reduction of client portfolio. This year, TICDES® decided to start an alternative software process initiative to attempt to establish an effective subcontracting process according to its characteristics.

6.1 Obtaining historical data.

Testing SSMM required selecting data from historical projects of TICDES®. It was possible to obtain data from one software project that failed in the external development delegation. The project PV0K_EXT1 was selected because it conserved all estimations and development data just before it was canceled. TICDES® does not have a formal process established to acquire external products or services, so our validation was focused on historical data. According to this, the supplier organization used a group of practitioners with university career formation to develop the required software products. A project manager and a four-member team were assigned to develop the project PV0K_EXT1. We discarded the “distance” factor as a cause of failure because both enterprises are located in the same city. Given the social and cultural factor of the development environment, we decided to use four students from last year as the

development team. These students have the same capabilities and knowledge as the original development team. We played the role of project manager (supplier) and TICDES® used the same project manager as in the original project (client). We satisfied the infrastructure requirement and assigned one PC per student; a server was used to set the SMMM assets and templates; and TICDES® used Internet connection to control the subcontracted project. Table 3 shows that TICDES®' historical data reflected the same problems determined in the previous work: inability to manage the subcontracted project.

Table 3. Historical data from project PVOK_EXT1

Objective	Estimated value TICDES®	Real value (at project cancellation)
Project planning (days)	180	300
Staff distribution	4	7
Role assignation	<1% of total time	> 5% of total time
Milestones reached (at first stage of project)	>30%	<10%
Predicted risks	5	15
Minutes of performed meetings	7 (one by life cycle phase)	2
Redacted agreements	7 (one by life cycle phase)	1 (development contract)
Modified documents (configuration management)	None	All TICDES® documentation was modified (project plan, roles, cost estimations, etc.)
Time for formation (days)	10	25
Time for tool identification (tools inventory for project)	20	35
SLA penalization	None	There is no SLA established
% of injected defects	< 5%	> 15%
Incidences in product quality	< 2% in relation to defects	Never used
Metrics for SLA	5 (all related to usability)	None
Effort deviation	< 15% of total time	> 32%

6.2 Establishing the evaluation objectives.

TICDES®' historical data enabled us to identify four specific issues: Lack of planning of subcontracted project; deficient monitoring of subcontracted project; nonexistence of SLA by each subcontracted project; poor quality in final product, if the project can be closed.

All these issues were coherent with the obtained results resumed in [11] with the application of a two-phase questionnaire. We launched a pilot project supported by SSMM called AP_PVOK (identical to PVOK_EXT1) and the preliminary results are showed in the following section.

6.3 Results

The pilot project AP_PV0K was implemented following our SSMM model to manage the software-subcontracting process. The student supplier development team followed the instructions of the TICDES® project manager and reported the project status using the SSMM assets.

Objective 1: Establish agile project planning in subcontracted projects.

SSMM introduced the Service Model to establish the project plan and IT plan. Table 4 depicts that the responsibilities matrix included in the Service Model considerably reduced the time spent in assigning roles and tasks for the project, and TICDES® was able to estimate the project risks with more accuracy.

Table 4. Reaching the objective 1 for pilot project

Objective	Estimated value TICDES®	Real value (at project cancella- tion)	Obtained value SSMM
Project planning (days)	180	300	160
Staff distribution	4	7	4
Role assignation	<1% of total time	> 5% of total time	<0.25 of total time
Predicted risks	5	15	7

Objective 2: Improve the monitoring process in subcontracted projects.

The contract asset of SSMM identified the inconsistencies between TICDES® and its original supplier, which may have resulted in the project failure after 10 months. The Service Model established monitoring meetings and used the SSMM templates to understand the agreements; the work was assigned between a Monitoring Committee and Maintenance Committee; and the criteria to establish monitoring milestones were established. The indicator of "Minutes of performed meetings" in Table 5 was incremented because SSMM established and maintained an Acceptation Criteria Plan, Final Acceptation Plan and Project Delivery Plan that TICDES® had never considered as formal documents.

Table 5. Reaching the objective 2 for pilot project

Objective	Estimated value TICDES®	Real value (at project cancella- tion)	Obtained value SSMM
Milestones reached (at first stage of project)	> 30%	< 10%	32%
Minutes of performed meetings	7 (one by life cycle phase)	2	10
Redacted agreements	7 (one by life cycle phase)	1 (development contract)	9
Modified documents (configuration man- agement)	None	All documents	0

Objective 3: Establish SLA to manage the subcontracting process.

The Metrics and SLA asset of SSMM established 24 metrics to manage the service level of products subcontracted by TICDES®. This activity increased the rate of confidence in the supplier work because it was constantly monitored, and TICDES® obtained what it really wanted. Using SSMM, a penalization percentage of 3% over total number of incidences was reflected in Table 6. We decide to show this penalization because the development team did not attend two of the monitoring meetings programmed by TICDES®.

Table 6. Reaching the objective 3 for pilot project

Objective	Estimated value TICDES®	Real value (at project cancella- tion)	Obtained value SSMM
SLA penalization	None	There is no SLA es- tablished	3%
Metrics for SLA	5 (all related to usability)	None	24

Objective 4: Establish and maintain the desired quality of product.

TICDES® used to apply two metrics to control the external products' quality: percentage of injected defects and the number of incidences in the final product. The person responsible for ensuring that these values do not exceed the established limit is the Software Quality Assurance Group. Unfortunately, as we explained before, it is too difficult for SMES to employ skilled personnel to perform this task. SSMM implemented 24 metrics and SLA from the Metrics and SLA asset through templates and activities to obtain better results.

The AP_PVOK project was closed after 4 months and, to this day, is working without quality problems. The project delivery was performed at the end of 2008 and obtained data was collected in February 2009. We are collecting more data to evaluate the effectiveness of the Service Model of SSMM.

Table 7. Reaching the objective 4 for pilot project

Objective	Estimated value TICDES®	Real value (at project cancella- tion)	Obtained value SSMM
% of injected defects	< 5%	> 15%	< 1%
Incidences in product quality	< 2% in relation to defects	Never used	1%

7 Conclusions

The externalization of software services is more frequently becoming an option among small enterprises as a solution for maintenance and new development of soft-

ware projects. The most important choice for enterprises is focused on choosing, in a formal way, who will be their supplier or technological partner. As part of this difficult decision, these organizations normally consider, among other aspects, the solid experience of their companion in this new journey. The model presented summarizes this experience in a document that enables the company, from the start, to take control of the service, making the right decision at the right moment and paying special attention to the relevant issues in each situation. We think that this model is the instrument to helping any small enterprises that provides software-subcontracting services, in managing it.

This alternative model represents the first step in this research. The next step is related to the validation of the model. For this purpose, the model is being experimented on 30 small-size enterprises through a project funded by the Spanish Ministry of Industry, Tourism and Trade.

Acknowledgement

This paper is sponsored by everis Consulting Foundation and Sun Microsystems companies through "Research Group of Software Process Improvement in Latin America".

References

1. Adler, P. *Practice and Process: The Socialization of Software Development*. Working Paper Series 03-12. Univ. Southern California (2003).
2. Bach, J. "The Immaturity of CMM" *American Programmer*, 7(9): 13-18 (September 1994).
3. Bernard, T., Gallagher, B., Bate, R. and Wilson, H. *CMMI Acquisition Module (CMMI-AM) Version 1.0*. Software Engineering Process Management. CMU/SEI-2004-TR-001 (2004).
4. Calvo-Manzano, J., Cuevas, G., Garcia, I., San Feliu, T., Serrano, A., Arboledas, F. and Ruiz de, F. "Requirements Management and Acquisition Management Experiences in Spanish Public Administrations" *International Journal of Knowledge Societies and Technologies*, 1(2): 116-121 (2007).
5. Clark, T. "Corporate Systems Management: An Overview and Research Perspective" *Communications of the ACM*, 35(2): 61-75 (1992).
6. CMMI Product Team. *CMMI for Development (CMMI-DEV, V1.2)*. CMU/SEI-2006 TR-008, Software Engineering Institute, Carnegie Mellon University.
7. COBIT 4th Edition Framework. COBIT Steering Committee and the IT Governance Institute (July 2005).
8. Dahane, M., Clementz, C. and Rezg, N. "Effects of extension of subcontracting on a production system in a joint maintenance and production context" *Computers & Industrial Engineering*, 58(1): 88-96 (2010).

9. Farbey, B. and Finkelstein, A. "Software acquisition: A business strategy analysis" *Proc. of the Fifth IEEE International Symposium on Requirements Engineering (RE'01)*, IEEE Computer Society, pp. 76-83 (2001).
10. Ferguson, E., Kussmaul, C., McCracken, D. and Robbert, MA. "Offshore Outsourcing: Current Conditions and Diagnosis", *Proc. of the ACM Special Interest Group on Computer Science Education (SIGCSE' 04)*, ACM Publications, pp. 330-331 (2004).
11. Garcia, I., Pacheco, C. and Sumano, P. "Use of Questionnaire-Based Appraisal to Improve the Software Acquisition Process in Small and Medium Enterprises" *Software Engineering Research, Management and Applications. Series: Studies in Computational Intelligence*, 150(14): 15-27. Springer-Verlag Berlin Heidelberg (2008).
12. Gilley, K., Greerb, C. and Rasheed, A. "Human resource outsourcing and organizational performance in manufacturing firms" *Journal of Business Research*, 57(3): 232-240 (2004).
13. Goldenson, D. and Fisher, M. *Improving the Acquisition of Software Intensive Systems*. Technical Report CMU/SEI-2000-TR-003. Software Engineering Institute, Carnegie Mellon University (2000).
14. Hietala, J., Kontio, J., Jokinen, J. and Pyysiäinen, J. "Challenges of Software Product Companies: Results of a National Survey in Finland" *Proc. of the 10th IEEE International Symposium on Software Metrics (METRICS'04)*, IEEE Computer Society, pp. 232-243 (2004).
15. ISO/IEC 15504-2:2003/Cor.1:2004 (E). Information Technology -Process Assessment -Part2. International Organization for Standardization: Geneva, 2004.
16. Kulpa, M. and Johnson, K. *Interpreting the CMMI: A Process Improvement Approach*. Auer Bach Publications. 2003.
17. Lee, J., Huynh, M., Ron, K. and Shih-Ming, P. "The Evolution of Outsourcing Research: What is the Next Issue?" *Proc. of the 33rd Hawaii International Conference on System Sciences*, Hawaii, USA (2000).
18. Lee, J., Huynh, M., Chi-Wai, R. and Shih-Ming, P. "IT Outsourcing Evolution: Past, Present, and Future" *Communications of the ACM*, 46(5): 84-89 (2003).
19. Office of the Government Commerce. *ITIL Lifecycle Publication Suite, Version 3: Continual Service Improvement, Service Operation, Service Strategy, Service Transition, Service Design*. Stationery Office Publisher (2007).
20. Rodgers, T. J. "The truth about outsourcing" *IEEE Design and Test of Computers*, 22(1): 12-13 (2005).
21. Software Acquisition Capability Maturity Model (SA-CMM) Version 1.03. Technical Report CMU/SEI-2002-TR-010 (March 2002).
22. Software Engineering Standards Committee of the IEEE Computer Society. IEEE Recommended Practice for Software Acquisition. IEEE STD 1062, 1998 Edition (Includes IEEE STD 1062-1993 and STD 1062A-1998) (1998).
23. Surmacz, J. *Take my Hosting Please*. Outsourcing Research Center Reports (June 2003).
24. Venkatraman, N. and Lohl, L. "Determinants of Information Technology Outsourcing: A Cross Sectional Analysis" *Journal of Management Information Systems*, 9(1): 7-24 (1992).

25. Venkatraman, N. and Lohl, L. "Diffusion of Information Technology Outsourcing: Influence Sources and the Kodak Effect" *Information Systems Research*, 3(4): 334-358 (1995).
26. Vivatanavorasin, C., Prompoon, N. and Surarerks, A. "A Process Model Design and Tool Development for Supplier Agreement Management of CMMI: Capability Level 2" *Proc. of the 13th Asia Pacific Software Engineering Conference (APSEC'06)*, IEEE Computer Society Press, pp.385-392 (2006).
27. Wang, Y. *Software Engineering Standards: Review and Perspectives*. World Scientific Publishing. 2002.
28. Weber, K., Araújo, E., Scalet, D., Andrade, E., Rocha, A. and Montoni, M. "MPS Model-Based Software Acquisition Process Improvement in Brazil", *Proc. of the Sixth International Conference on the Quality of Information and Communications Technology*, IEEE Computer Society, pp. 110-119 (2007).
29. Wong, S. "Software Acquisition Management Experience Learnt in a Multi Discipline and Multi Contract Project Environment" *Proc. of the First Asia-Pacific Conference on Quality Software*, IEEE Computer Society, pp. 239-247 (2000).

Supporting the Management Process of Software Process Improvement Initiatives based on NMX-I-059/02-NYCE-2005

Garcia, I. and Cruz, D.

Postgraduate Department
Technological University of the Mixtec Region, Mexico
{ivan@mixteco.utm.mx, dago@mixteco.utm.mx}

Abstract. Nowadays there are models and standards which attempt to introduce quality in the enterprises' software development process with the objective to introduce high quality levels in the produced software. The NMX-I-059/02-NYCE-2005 standard (also known as MoProSoft) is focused on small and medium software enterprises, or small groups of software development within a larger organization, with the aim of promoting the standardization of an effective process in the software industry. Mexican enterprises now have a software standard that enables them to achieve a high level of quality in the software that they produce. However, the adoption of any standard is not an easy task. This paper aims to show that the development and implementation of a RIA-based tool that could support improvement initiatives, therefore strengthening the standard adoption.

Keywords: Software process improvement, effort and adoption time, MoProSoft, small software enterprises, process assessment and improvement.

1 Introduction

Software has arisen as a fundamental pillar in the evolution of computational products and services. In the last two decades, software has changed from a "specific problem solution" to an autonomous industry [11]. However, this change keeps the old same problems since the "software crisis" in 1969. Nowadays, the quantity and quality demands dominate the market. According to the last report of Standish Group Inc. [38]:

- Software is (almost) always delivered out of the initial planning,
- Software is more expensive than the original cost,
- Software has a different functionality.

Pressman says that: "*the majority of software crisis causes have their origins in myths and theories that arose in the early years of Software Engineering*". This origin makes the myths more dangerous; but the truth is that they do not look like myths any more. Figure 1 shows that a recent study [39] establishes that software does not accomplish the original requirements because: 45% of software exceeds the cost, 63% of software exceeds the planned schedule, and software fulfills the 67% of the required functionality.

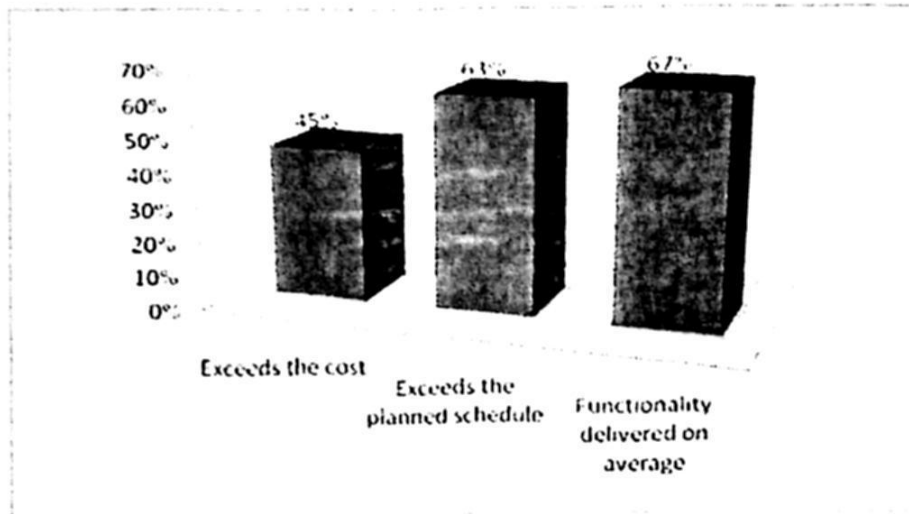


Fig. 1. Principal causes of failure in software projects

Software industry has tried to increase productivity and quality adopting alternative methodologies and technologies, but it has been recognized that the main problem is related to the incapacity to manage the software process [27] [17]. In this way, enterprises have been changed from technological-based solutions to process-based solutions. From the beginning of the 90's, industry and researchers interested in Software Engineering have been expressing special interest in Software Process Improvement (SPI) [21]. An indicator of this is the increasing number of international initiatives related to SPI, such as CMMI-DEV v1.2 [35], ISO/IEC 15504:2004 [14], SPICE [13], and ISO/IEC 12207:2004 [15].

In addition, many methods for evaluating improvements in organizations, such as SCAMPI [19], ISO/IEC 15504:2004 and CBA-IP1 [4], and improvement models such as IDEAL [18] have been developed. This interest in software improvement in large enterprises is now being extended to small enterprises. However, the problem is the high implementation cost, independent of the size of the company [5]. Because models have been developed for large enterprises, only a few small software Enterprises (SE) are aware of them. In Mexico, in April 2007 there were almost 1,500 SE that accounted for 99.87 % of all companies in this category. Due to this, the Software Industry Process Model (MoProSoft) [23] and its process-assessment method (EvalProSoft) were defined [24].

Since the approval of MoProSoft and EvalProSoft as the Mexican standard NMX-I-059-NYCE-2005, the interest of the SE on acquiring this certification has grown. Besides the certification, as a result of the correlation between the process quality and the obtained product quality, the Mexican SE have gained impulse to improve their software processes as a strategy to assure the quality of their software products [6]. However, besides the certification desire, the SE has the problem to adopt quickly and efficiently the standard NMX-I-059-NYCE-2005 without loss resources and time to develop new projects.

This paper presents the experiences on adopting a RIA-based tool as support to facilitate the certification process by four small software companies located in the states

of Tlaxcala, Puebla and Oaxaca. The document is structured as follows: Section 2 is an introduction to the MoProSoft model and Section 3 discusses related work. Section 4 describes the architecture and development of RIA-based tool. Section 5 describes the methodology used and presents the characteristics of the case study. Section 6 describes the initial scenario prior to the tool support and presents results and lessons learned. Finally, conclusions are shown in Section 7.

2 The MoProSoft Model

In 2006 the Mexican government, through the National Plan for Development, established the objectives to increase and extend the country's competitiveness using the information and technology. One of the crucial issues to address was to "promote the IT industry development" Thus, the Economics Secretary and enterprise organisms designed the Software Industry Development Program (PROSOFT) to promote the software industry and extend the IT market in Mexico. According to the government and industry requirements, the selection process of one model that would be adopted by the Mexican software industry it would be based on five criteria [25]. The model should be:

- Adequate for SE with a low maturity level,
- Affordable cost of adoption and evaluation,
- Acceptable as national norm,
- Specific for software enterprises,
- Defined as a set of processes based on international practices.

According to Oktaba [25] no one of the analyzed models (ISO 9000:2000 [32], CMM/CMMI [34], ISO/IEC 12207 [15], and ISO/IEC 15504 [14]) completely fulfilled the established criteria. The new proposed model, MoProSoft, is developed taking into account the better practices of models as CMMI-SW [34], ISO 9000:2000, PMBoK [29], among others. This model provides a new process structure, new elements to document the process, a more precise relation among processes, and an explicit mechanism for SPI [25] [26]. MoProSoft is conformed by the Software Industry Process-Assessment Method (EvalProSoft) which is based on ISO/IEC 15504 [Part 2] recommendations. Figure 2 shows the MoProSoft' categories emphasizing that, unlike other reference models, has the Top Management category, whose main objective is to organize the activities of the SE executives by means of introducing modern management and Software Engineering practices.

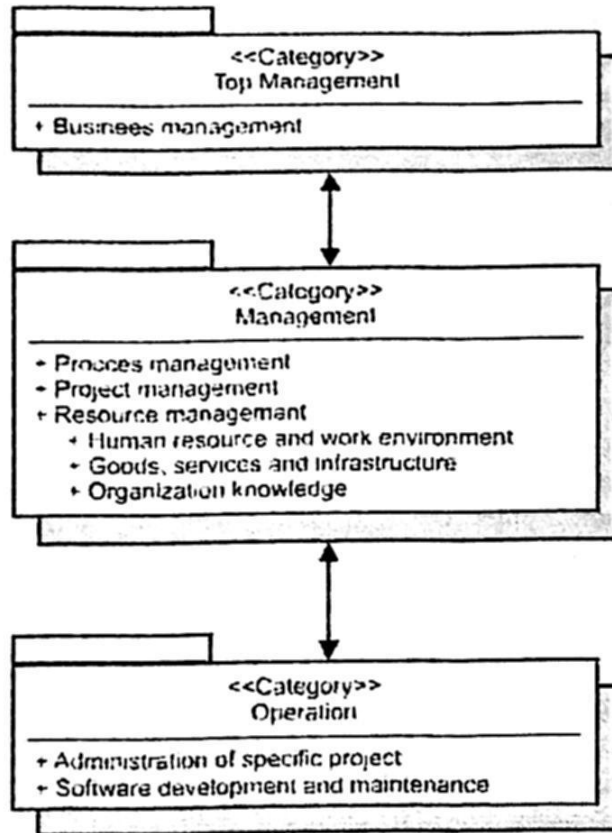


Fig. 2. MoProSoft structure

To carry out an initiative of SPI within an organization, it is necessary to involve: a model that guides the improvement, a method for assessing a process, and a reference model to follow. The capability levels achieved and their process attributes are located over a scale of six levels, where the level 0 is associated with the lowest capability level, indicating that the purpose of the process is not reached. Table 1 shows the levels which are used to determine whether a process has reached a capability level.

Table 1. MoProSoft capability levels

Process attributes	
Level 0 Incomplete	It means that the process has not reached the expected objectives.
Level 1 Realized	The implemented process achieves its purpose and obtains defined results.
Level 2 Managed	The realized process is implemented in an administrative way, and the work products are properly established, controlled and maintained.
Level 3 Established	The process is managed and implemented through defined objectives, which are capable of achieving the desired outcome.
Level 4 Predictable	The established process operates within certain limits to achieve their results.
Level 5 Optimized	The predictable process is continuously improved in order to achieve current and future applicable business goals.

At the end of 2009, 171 organizations have been assessed under the standard NMX-I-059/02-NYCE-2005. From these, 93% has level 1 and 5% level 2 [40]. But, Mexican industry should accelerate the standard adoption, and according to Young et. al [43] a way to do it is introducing software tools to process improvement.

3 Related Work

Rapid assessments are a success factor for the SPI efforts of any organization because they can be frequently applied, with minimal effort-time and resources. These assessments provide information to enterprises about the impact of improvement actions, introduced through a SPI effort, in processes. In literature there is some research related to rapid assessments: SEAL OQ [12], PIASS [22], SPIS [8], KMT [1], SPQA.Web [28], and some commercial tools created by specialized companies: CMMI v1.2 Browser [42], CMM-Quest [33], Appraisal Wizard [2], SPICE 1-2-1 [36], IME Toolkit [16], MKS Integrity Suite [20], and Stages for CMMI [37].

The standard NMX-I-059/02-NYCE is relatively new since its creation at 2005. There are some researches which try to accelerate and support the SPI initiatives focused on MoProSoft; however, none of them could be considered a SPI tool.

Research by [41], for example, provides a tool for accelerating the adoption of MoProSoft and is supported by the AceleraProSoft¹ project. The main functionality of Kuali tool is that it provides a mechanism to manage and control all documents that result from implementing the MoProSoft processes. But, it does not provide SPI support to assess or adopt efficiently the model.

Caballero [3] provides a document management tool, named MDM, that supports enterprises to document the MoProSoft processes; specifically business management, process management, and administration of specific projects. MDM was developed to provide support via Web and enables the creation of templates for each process. However, MDM is not an SPI tool.

The Guiding and Monitoring Tool for Automation of MoProSoft (Assistant HIM) [44] provides the support to adapt and monitor the model through a Web environment. Assistant HIM is an electronic guide that shows processes, activities, product, and roles according to MoProSoft. The Web tool uses a knowledge base (developed with the Resource Description Framework) that generates the needed information for model in XML files. Nevertheless, Assistant HIM assumes that the SPI effort was already conducted and provides only a guide to "easily" adopt the model, but it is not a SPI tool too.

Reyes et. al [30] provides an Instrument of Self-Assessment for Diagnosing the Software Process using MoProSoft. This tool is focused on the process management process and assesses the organizations to identify improvement areas. The collected information is quantitative and qualitative and correspond with a Likert scale [31] that

¹ AceleraProSoft is an initiative developed by the Mexican Economics Secretary, together with Microsoft, Visionaria and Amity. This initiative attempts to guide enterprises to increase their sells from short and medium term. Besides, it focused its efforts to strengthen the planning, operation, sells management, and marketing research.

represents the MoProSoft' capability levels showed in Table 1. This tool could be considered as a SPI tool because it assesses the software process and identifies an organization's strengths and weaknesses; however, it only assess and does not provide an improvement plan nor monitoring the improvement activities.

Summarizing, Table 2 shows a comparison among the analyzed tools according to their function and operation perspectives with the aim of highlighting some gaps and obtain an initial benchmark.

As we can see, the available tools could be used to rapidly adopt the standard NMX-I-059/02-NYCE-2005, however there is not information about their success in real small environments. That is, there is a limit to being able to provide more helpful and diverse support for establishing SPI initiatives using MoProSoft and managing the whole process of SPI.

We provided an alternative way to reduce the effort of adoption using a Rich Internet Application (RIA) tool. In addition, the proposed tool provides not only a fundamental process assessment and improvement features, but also an improvement plan generation and remote online assessment and self-evaluation respectively depending on organizations.

Table 2. MoProSoft' tools features

Criteria	MoProSoft tools			
	Kuali	MD M	Assistant HIM	Self- Assessment
Supports SPI initiatives	No	No	No	No
Assesses and generates improvement plans	No	No	No	Only assess
Obtains a snapshot of current process	No	No	No	Yes
Covers all model processes	No	No	Yes	No
Developed for Web environment	No	Yes	Yes	Yes
Measures and monitoring the SPI initiative	No	No	No	No
Applied on SE	No	No	No	No

4 A RIA-based tool for SPI

The SelfVation (SELF-eValuATION) tool provides support to easily adopt the MoProSoft model. SelfVation is a RIA-based tool. RIAs are web applications that have most of the characteristics of desktop applications, typically delivered either by way of a standard based web browser, via a browser plug-in, or independently via sandboxes or virtual machines. Examples of RIA frameworks include Ajax, Curl, GWT, Adobe Flash/Adobe Flex/AIR, Java/JavaFX, Mozilla's XUL and Microsoft Silverlight.

We are trying to provide a RIA-based tool that establishes an iterative approach to process improvement using the standard NMX-I-059/02-NYCE-2005 and which a small organization could adopt. Following this approach, the focus of the first step

would be to understand what exists in the organization and determine what causes significant problems. Then solutions could be devised in the action plan and evaluated in pilot studies or even controlled experiments.

SelfVation' architecture is based on three layers separating the configuration settings, the assessment and improvement mechanisms, and information retrieval (see Figure 3).

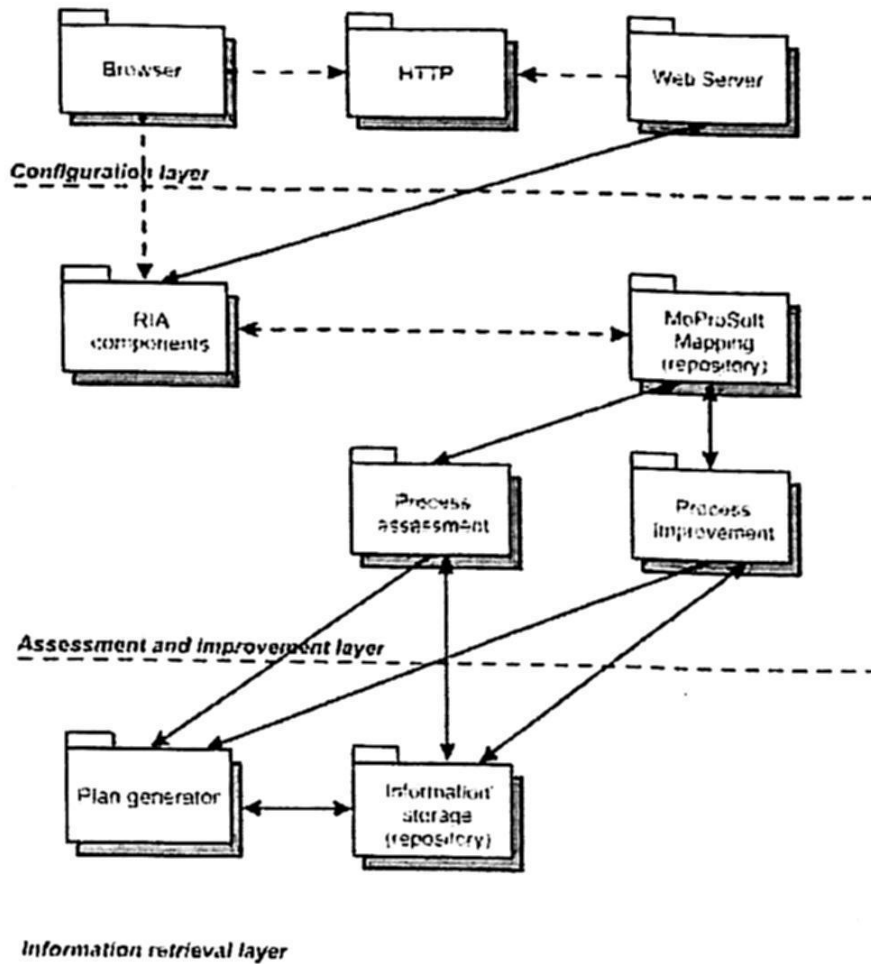


Fig. 3. The architecture of SelfVation

This kind of architecture makes possible that through the configuration layer, all SEs take control over the complete SPI initiative and obtain, at the same time, customized results; all the business characteristics are evaluated with the assessment and improvement layer, establishing a real mapping of current software processes. The information retrieval layer manages all the generated information and uses the plan generator to send the complete information to clients. This architecture provides a light application for any client, and offers advantages related with implementation and management capabilities in a flexible computational program.

Figure 4 illustrates that the SelfVation strategy begins with an evaluation that is performed in two ways: *process modeling* and *questionnaires*. Both results are mapped against the standard NMX-I-059/02-NYCE-2005 and results are analyzed by an internal manager who generates and improvement plan. SelfVation covers the Top

Management category managing the top management level information, and the Management and Operation categories managing the practices conducted by the project managers.

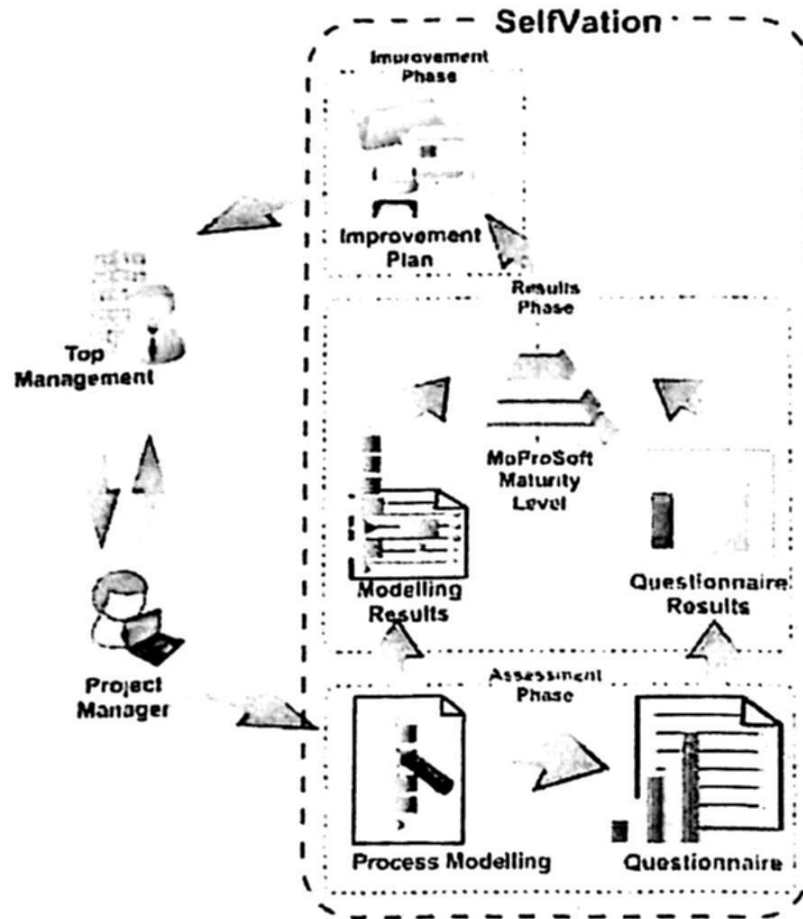


Fig. 4. SelfVation modules

SelfVation provides three types of user interface modules: (1) the assessment phase, (2) the results phase, and (3) the improvement phase.

4.1 The assessment phase.

This phase guides the project managers to obtain all knowledge from their daily labor. Previously, top management had selected for MoProSoft' processes to evaluate. This interface uses a refined version of the Liker scale and establishes 'levels of performance' from the two-phase questionnaire proposed in [10]. Table 3 shows that each answer of Liker scale corresponds with an established level of performance to determine the percentage in which each practice is performed.

Table 3. SelfVation' perform levels

Possible answer	Perform level	Description
<i>Always</i>	4	The activity is documented and established in the organization (between 75% and 100% of the time).
<i>Usually</i>	3	The activity is established in the organization but rarely documented (between 50% and 74% of the time).
<i>Sometimes</i>	2	The activity is weakly established in the organization (between 25% and 49% of the time).
<i>Rarely</i>	1	The activity is rarely performed in the organization (between 1% and 24% of the time).
<i>Never</i>	0	The activity is not performed in the organization.

Giving a specific weight to each response will enable us to easily analyze the results of the evaluation and to identify which practices are common within the whole organization and which ones are not performed at all. The second part of assessment is related to process modeling. SelfVation uses the Process Change Methodology [7] to describe the current process based on MoProSoft and mapping it against an ideal process according to the model. None of existing tools had ever been used this kind of assessment. SelfVation introduces the use of graphical notation to assess or modify current and pilot processes.

4.2 The result phase.

This phase, which allows members of top management to obtain any information on the SPI cycle at any time; they can control the performance of its project managers through the assessment phase and obtain the final results and graphics derived from the entire process. The mapping process presents a categorized level of performance, in accordance with the assessed process. The project manager can meet its own level of performance. This phase just provides performance level results to project managers; the entire results of the organization can only be reviewed by top management through a reports generator.

4.3 The improvement phase.

Improvement phase takes the results obtained in previous phase and provides the mapping with the knowledge base that contains the MoProSoft activities. The improvement guide is offered in a RIA improvement performance. This means, that the improvement mechanism of SelfVation can avoid the latency of round-trips to the server by processing locally on the client and are often a lot faster. Offloading work to the clients can also improve server performance. Conversely, the resource requirements can be prohibitive for small, embedded and mobile devices.

5 Methodology

On the graduate program of Masters in Computer Science of the Technological University of the Mixtec Region (UTM), a research project was conducted to obtain data that contribute to a successful implementation of the SelfVation tool. The main tasks of the methodology for this project were:

1. Identification of small software organizations which want to assume an improvement commitment,
2. Using SelfVation to assess the organizations' actual situation,
3. Using SelfVation to improve the weakness detected in step 2.
4. Implementing the mandatory processes awareness through the tool.
5. Performing and official assessment of compliance with MoProSoft capability level.

In order to choose the right SEs, we decided to experiment with three kinds of enterprises: ones which know and uses the standard, ones which only know the standard, and those which do not know the standard. Since our main objective is to reduce the effort-time to adopt the standard NMX-I-059/02-NYCE-2005, we believed that one of the crucial factors is that SelfVation provides the same guide to those enterprises which really implement the model and the others who do not have previous knowledge. Table 4 shows the characteristics of participating small enterprises.

Table 4. Profiles of the participating SEs

SE	Activity	Size	Experience
1	Software development and maintenance	10	The standard is implemented
2	Software development and maintenance	20	The standard is known, but not implemented
3	Development of software solutions for small enterprises	15	The standard is not known
4	Hardware & software services and integral solutions	8	The standard is known, but not implemented

A case study is commonly employed as an empirical research strategy in information systems field, often used for describing relationships within organizational settings [9]. In the context of our research, we employ a case study as the method to provide an organizational context for the application of SelfVation for identifying SPI factors and their implementation and deployment to four particular small software companies.

6 A case study: adopting SPI initiatives in four Mexican small software enterprises

We designed a controlled experiment with four small enterprises in which we focus our effort in a semiformal version of the EvalProSoft approach. An assessment team and four project managers were chosen, besides we separated the four organizations into three categories according to their experience with the standard. The selected project managers are professional who know the enterprise's culture and the way that the development projects are conducted. The project managers had received the standard NMX-I-059/02-NYCE-2005 basic training. In a similar manner the project managers had answered a structured questionnaire related to standard' categories. Figure 5 illustrates the assessment phase using questionnaires; the answers are agreed with the way that the enterprise works by each phase of the development lifecycle that the standard establishes (left side of the Figure 5).

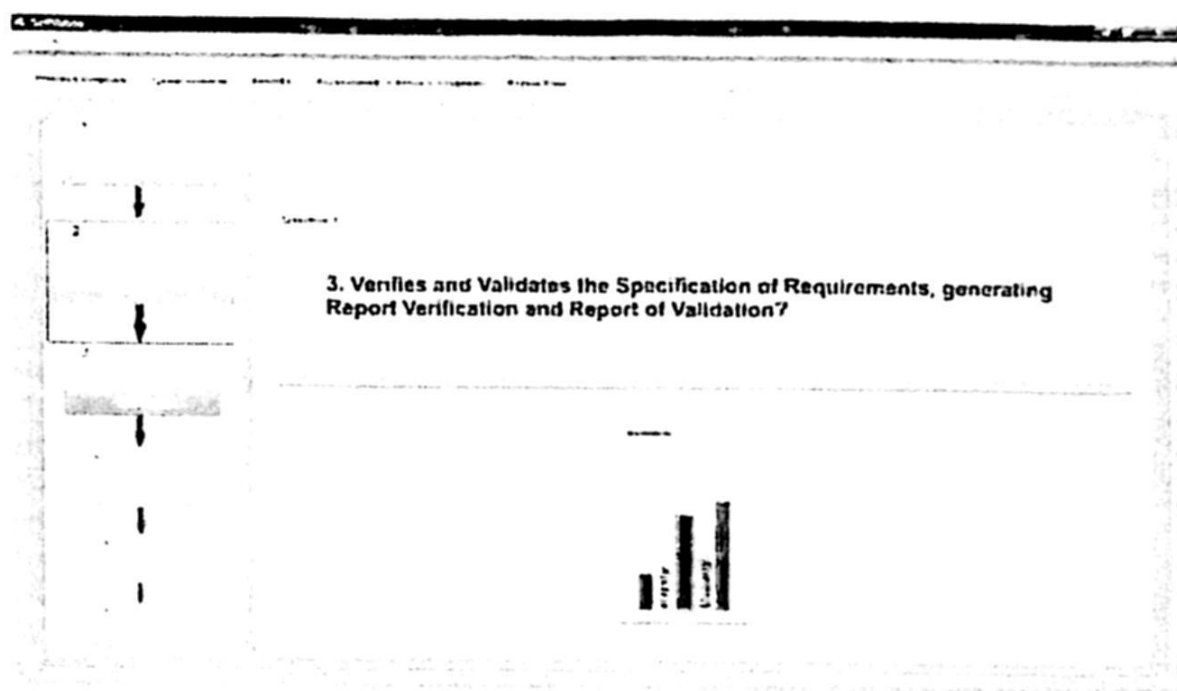


Fig. 5. Assessment phase

Once the questionnaires were completed we proceed to assess enterprises using the modeling process evaluation. This RIA component provides an easy interface to capture the essence of SE' development process (see Figure 6).

These two assessment components enable us to identify the strong and weak points in order to show the current situation of the four enterprises according to the standard NMX-I-059/02-NYCE-2005, so that it is indicated where they focus their efforts to raise the quality of the software that they develop and maintain. Figure 7 shows the coverage of the standard for enterprise SE1.

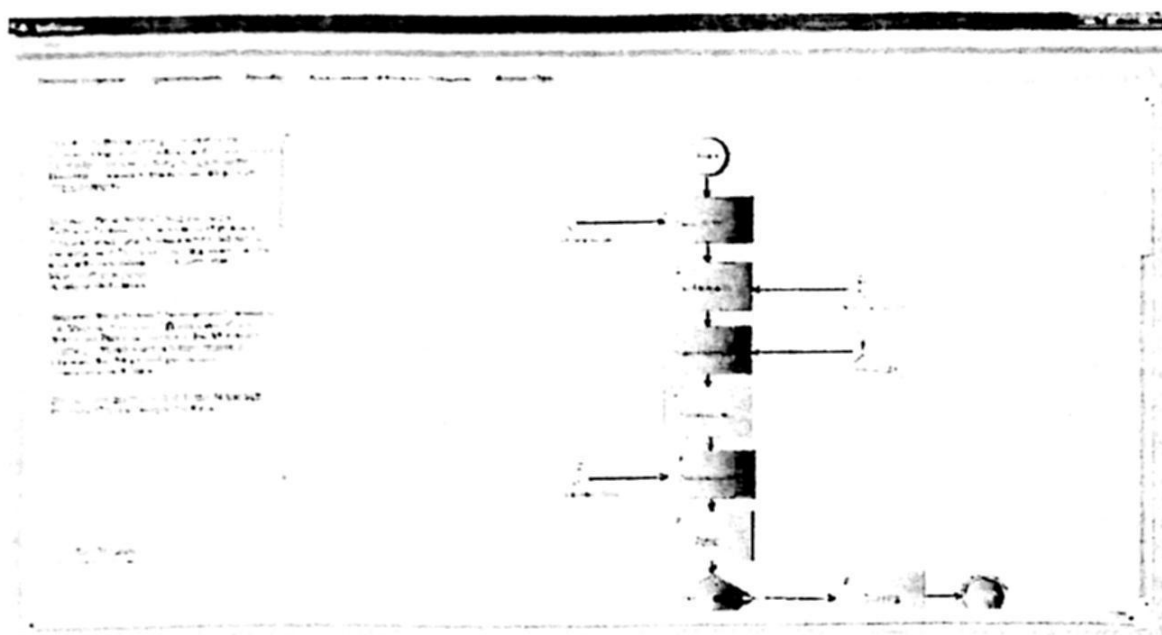


Fig. 6. Process modeling assessment

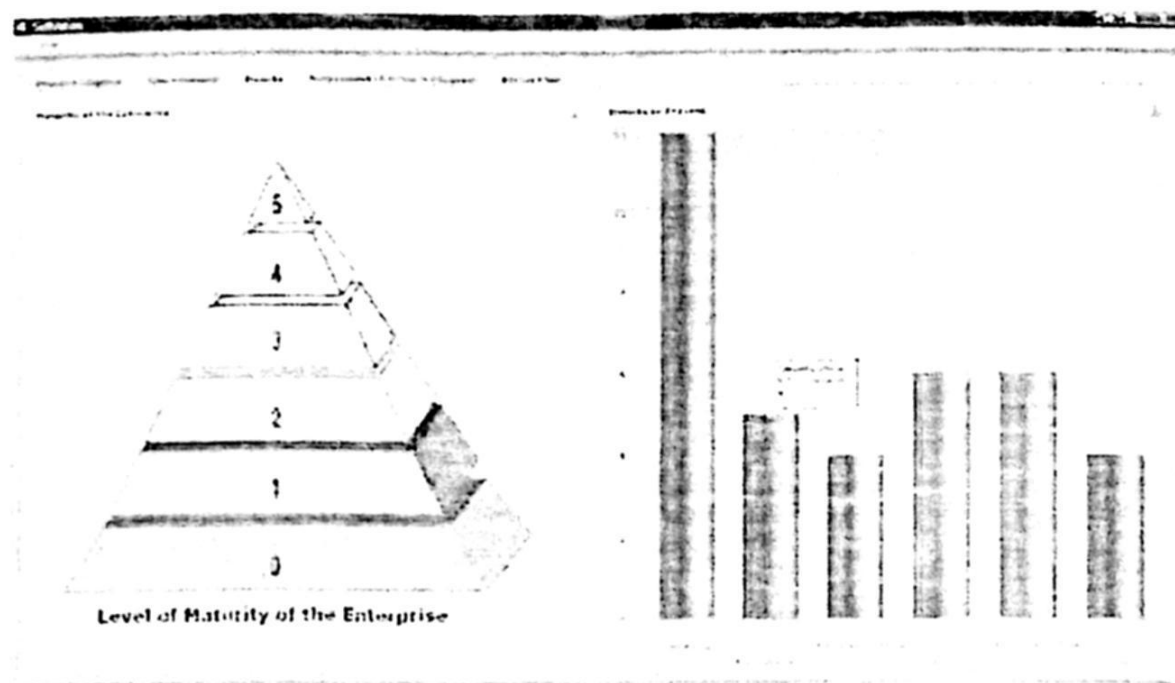


Fig. 7. Results phase

These mapped results are used by the reports generator module and create an improvement plan according the current situation for every organization. The improvement plan of Figure 8 shows different domains (practices, products, inputs, outputs, measurement criteria, entry criteria, exit criteria) which enable small companies to establish the corrective actions indicated in the final report.

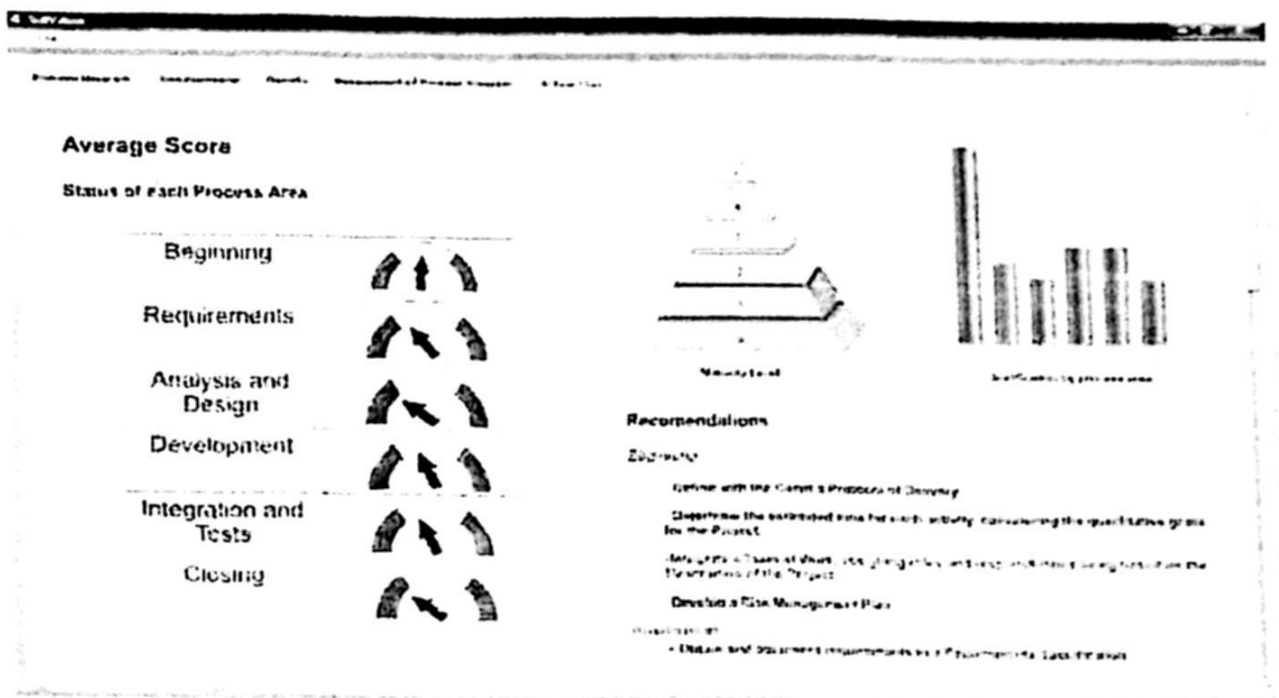


Fig. 8. Improvement phase

6.1 Experimental results and learned lessons.

After implementing the assessment and improvement phase we obtain information about the practices and effort performed by the four SE. As we said before, SE1 has an official implementation of standard NMX-I-059/02-NYCE-2005 and provides us a measure of contrast. All enterprises were assessed in Level 1 of MoProSoft (a realized level) and Figure 9 illustrates the obtained results.

We believe there exists a relation between the enterprise size and the effectiveness for standard adoption. We do not try to affirm that there exists a proportional relation between both factors, but it is possible that small teams are more organized and make a rapid adoption of standard. Figure 9 shows that SE1 obtained a 61.5% of coverage in Level 1 that corresponds with its real conformance level. SE2 and SE4 obtained similar coverage results, approximately a 37%, which are insufficient to achieve Level 1 (according to the assessment mechanism over the 60%). SE3 obtain a lower coverage ratio of 14.8% that locate it in an *ad-hoc* or chaotic situation. These results were obtained from a launch of an improvement initiative using SelfVision.

On the other hand, the coverage by each activity of standard NMX-I-059/02-NYCE-2005 was obtained. Figure 10 shows that the best coverage is obtained for the Requirements activity (60% of average and 17% of standard deviation) and indicates that the requirements are used as a basis to develop a plan and construct the software product. The worst coverage is obtained for Integration and Test activity (22% of an average and 17% of standard deviation) because small companies do not use formal methods of product validation/verification.

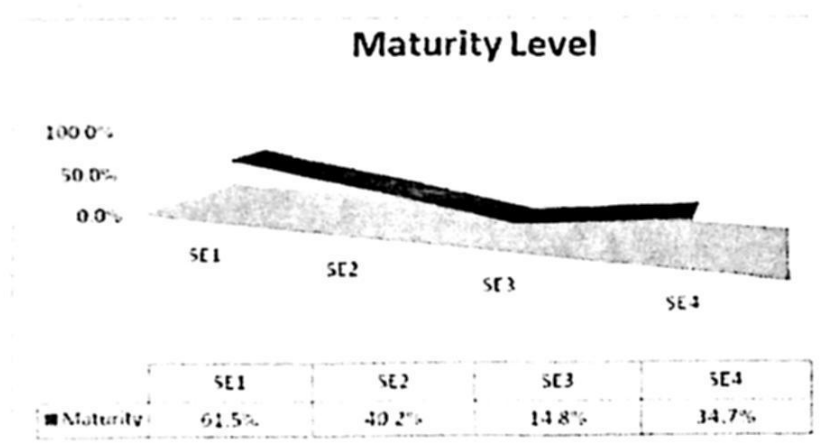


Fig. 9. Maturity level per enterprise

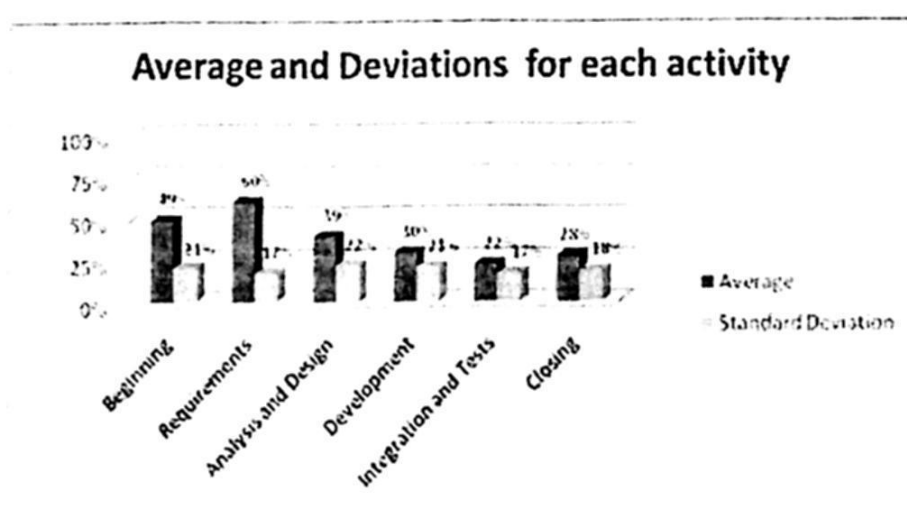


Fig. 10. Coverage per activity

The purpose of this experiment was to demonstrate the ease and usefulness of SelfVation as an SPI tool for SE, which also reduces the costs of undertaking an SPI initiative, considering the particular conditions of these organizations. The initial assessment with SelfVation was conducted to establish the base line capabilities of the enterprises processes according to the standard NMX-I-059/02-NYCE-2005. The result was "poor" –between 0 and 1 (according to MoProSoft levels). During the following 3 months SelfVation coached the enterprises on SPI tailoring and adoption through improvement plans.

Finally, we applied the second assessment with SelfVation to each enterprise; all four enterprises achieved a 1.00 average increase in the Maturity Level of two areas: Administration of Specific Project and Software Development and Maintenance. This increased level is obtained from the initial process coverage over 100%. For example, SE1 has an initial coverage of 61.5% when the first assessment was conducted. A reassessment after using SelfVation showed an increase in processes of 0.28 (in first iteration). Therefore the process has an improved level of $61.5\% + 61.5\% \times 0.28$.

Table 5. Improvement and effort data by enterprise

	Enterprises			
	SE1	SE2	SE3	SE4
Total effort (hours)	287	484	554	220
Total effort SelfVation	243	400	495	185
Effort per person (hours)	28.7	24.2	36.9	27.5
Effort with SelfVation	14.2	11.3	17.9	20.2
Improvement average	0.28	0.86	1.43	1.00

An official assessment in SE1 demonstrates that the SelfVation' results coincide with obtained data; but we obtained this information with less effort, it means we are capable to reduce the time of adoption.

These results show the applicability of our tool; however we need to experiment with more than four SEs. We are designing a new experiment that involves 15 small companies from different cities in the Mexican Republic. The lessons learned will be discussed in future research.

7 Conclusions

The implementation success when enterprises try to adopt an improvement initiative depends on the organization's top management commitment. Small companies are not the exception, but this research enables to identify that SE does not understand the benefits that this improvement process would have in the company. To initiate the improvement program, involvement and experience of the key personnel could be the key factors that contribute to strengthen the improvement initiative. But, the standard NMX-I-059/02-NYCE-2005 is relatively new in Mexico and small companies have no experience in SPI.

This research, therefore, has developed an instrument to define and implement SPI initiatives to improve the current status of SE practices using the standard NMX-I-059/02-NYCE-2005 as reference model. Its purpose was to investigate its feasibility in SE and to influence the direction of future research. One limitation of this study is the generalization of its findings based on the limited amount of data collected and analyzed relative to the number of small organizations. This suggests that this qualitative study will be augmented by quantitative studies to strengthen the data supporting the need and applicability of SelfVation to the Mexican small organization community.

At the moment of this report, the selected SEs have decided to improve their processes in order to achieve a level 2 certification. We are collaborating through the implementation of an SPI iterative cycle using SelfVation.

References

1. Alagarsamy, K., Justus, S. and Iyakutti, K. "On the Implementation of a Knowledge Management Tool for SPI" Proc. of the International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 07. IEEE Computer Society, pp. 48-55. 2007.
2. Appraisal Wizard, Formal or Informal Appraisal Tool, 2009. Available at: <http://www.isd-inc.com>
3. Caballero-De la Villa, D. "Manejador de Documentos de MoProSoft". Universidad de las Americas Puebla. Mayo, 2005 (in Spanish).
4. Dunaway, D.K., Masters, S. CMM-Based Appraisal for Internal Process Improvement (CBA IPI). Method Description. Technical Report CMU/SEI-96-TR-007. Carnegie Mellon University, Software Engineering Institute, Pittsburgh. 1996.
5. Dybå, T. "Factors of software process improvement success in small and large organizations: an empirical study in the Scandinavian context" ACM SIGSOFT Software Engineering Notes, 28(5): 148-157. September 2003.
6. Flores, B., Astorga, M., Olguín, J. and Andrade, M. C. "Experiences on the Implementation of MoProSoft and Assessment of Processes under the NMX-I-059/02-NYCE- 2005 Standard in a Small Software Development Enterprise" Proc. of the 2008 Mexican International Conference on Computer Science, IEEE Computer Society, pp. 323-328, 2008.
7. Fowler, P.; Middlecoat, B.; & Yo, S. Lessons Learned Collaborating on a Process for SPI at Xerox (CMU/SEI-99-TR-006, ADA373332). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1999.
8. Fukuyama, S., Miyamura, S., Takagi, H. and Tanaka, R. "Validating the ISO/IEC 15504 Measure of Software Requirements Analysis Process Capability", IEICE Transactions on Information and Systems, E83-D(4): 747-756, 2000.
9. Galliers, R. Information Systems Research. Issues, Methods and Practical Guideline. Alfred Waller Ltd. Chippenham, Wiltshire. England. 1992.
10. Garcia, I., Calvo-Manzano, J., Cuevas, G. and San Feliu, T. "Determining Practice Achievement in Project Management Using a Two-Phase Questionnaire on Small and Medium Enterprises" Proc. of the European Systems and Software Process Improvement and Innovation Conference, EUROSPI 2007, LNCS 4764, pp. 46-58, Springer-Verlag Berlin Heidelberg. 2007.
11. Girba, T. and Ducasse, S. "Modeling history to analyze software evolution," Journal of Software Maintenance: Research and Practice (JSME), vol. 18, pp. 207-236, 2006.
12. Him Lok, R. and Walker, A. J. "Automated Tool Support for an Emerging International Software Process Assessment Standard" Proc. of the Third International Software Engineering Standards Symposium; Emerging International Standard, ISESS 1997. IEEE Computer Society, pp. 25-35. 1997.
13. ISO/IEC TR 15504:1998(E): Information Technology – Software Process Assessments. Parts 1-9. International Organization for Standardization: Geneva. 1998.
14. ISO/IEC 15504-2:2003/Cor.1:2004(E): Information Technology – Process Assessment – Part 2: Performing an Assessment. International Organization for Standardization: Geneva. 2004.

15. ISO/IEC 12207:2002/FDAM 2: Information Technology – Software Life Cycle Processes. International Organization for Standardization: Geneva. 2004.
16. IME Toolkit, 2009. Available at: http://www.man-info-systems.com/index_files/FreeTools.htm
17. Johnson, J. *My Life Is Failure: 100 Things You Should Know to Be a Better Project Leader*. Standish Group International Publisher. 2006.
18. McFeeley, B. "IDEAL: A User's Guide for Software Process Improvement" CMU/SEI-96-HB-001, Software Engineering Institute, Carnegie Mellon University. 1996.
19. Members of the Assessment Method Integrated Team. Standard CMMI® Appraisal Method for Process Improvement (SCAMPI), Version 1.1. CMU/SEI-2001-HB-001. Software Engineering Institute, Carnegie Mellon University. Pittsburgh, PA. 2006.
20. MKS Tool, 2009. Available at: http://www.mks.com/process_improvement
21. Moe, N. B. and Dybå, T. "Improving by Involving: A Case Study in a Small Software Company" *Proc. of the European Software Process Improvement Conference 2006 (EUROSPI 2006)*, LNCS 4257, pp. 159-170, 2006.
22. Nakakoji, K. "PIASS: Process-Improvement Activity Support System" Technical Report SRA-SEL-97081, Software Engineering Lab., SRA Inc., Tokyo, Japan. 1997.
23. NMX-I-059/01-NYCE-2005. Information Technology – Software – Process and Assessment Model to Software Development and Maintain – Part 02: Processes requirements (MoProSoft). NMX-NYCE. 2007.
24. NMX-I-059/04-NYCE-2005. Information Technology – Software – Process and Assessment Model to Software Development and Maintain – Part 04: Guidelines for processes assessment (EvalProSoft). NMX-NYCE. 2007.
25. Oktaba, H. "MoProSoft: A Software Process Model for Small Enterprises". *Proc. of the First International Research Workshop for Process Improvement in Small Settings*, Software Engineering Institute, Carnegie Mellon University. Special Report CMU/SEI-2006-SR-001, pp. 93-101, 2006.
26. Oktaba, H., García, F., Piattini, M., Ruiz, F., Pino, F. and Alquicira, C. "Software Process Improvement: The Competisoft Project" *Computer*, 40(10): 21-28. 2007.
27. Oktaba, H. and Piattini, M. *Software Process Improvement for Small and Medium Enterprises: Techniques and Case Studies*. Information Science Reference Publisher. 2008.
28. Pino, Francisco J., García, F. & Piattini, M. "Herramienta de Soporte a la Valoración Rápida de Procesos Software". *IEEE Latin America Transactions*, 5(4). July, 2007 (in Spanish).
29. Project Management Institute. *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*. Project Management Institute. 2004.
30. Reyes, P. Y., Margarin, M. L., Álvarez, F. and Muñoz, J. "Diseño de un Instrumento de Auto-evaluación para Diagnosticar el Estatus de las Organizaciones en México con Respecto al Modelo ProSoft: Proceso de Gestión de Procesos de la Categoría Gestión". Universidad Politécnica de Aguascalientes. 2008 (in Spanish).
31. Russell, C. and Bobko, P. "Moderated Regression Analysis and Likert Scales too Coarse Comfort" *Journal of Applied Psychology*, 77(3): 336-342. 1992.

32. Russell, S. "ISO 9000:2000 and the EFQM Excellence Model: competition or co-operation?" *Total Quality Management*, vol. 11, no. 4/5&6, pp. 657-665, 2000.
33. Self Assessment Tool CMM-Quest, 2009. Available at: <http://www.cmm-quest.com/>
34. Software Engineering Institute. CMMI for Systems Engineering, Software Engineering, Integrated Product and Process Development, and Supplier Sourcing (CMMI-SE/SW/IPPD/SS, V1.1). Continuous Representation. CMU/SEI-2002-TR-011, Software Engineering Institute, Carnegie Mellon University. 2002.
35. Software Engineering Institute. CMMI for Development (CMMI-DEV, V1.2) CMU/SEI-2006 TR-008, Software Engineering Institute, Carnegie Mellon University. 2006.
36. SPICE 1-2-1, 2009. Available at: <http://www.synspace.com/tools.html>
37. Stages for CMMI, 2009. Available at: <http://www.methodpark.com/en/products/stages-special-editions/stages-for-cmmi-process-management-process-asset-library/>
38. Standish Group International, Inc. CHAOS Report 2007: The Laws of CHAOS. 2007.
39. Standish Group International, Inc. CHAOS Summary 2009. 2009.
40. TI-NYCE Verification unity. Reports list. Available at: <http://www.nyce.org.mx/dictamenes.htm> [Online] November, 2009.
41. Strevel, C. "Kuali: Herramienta Auxiliar para implementación de MoProSoft". DevDays, Intellect. 2005 (in Spanish).
42. Wibas CMMI Browser, 2009. Available at: http://www.cmmi.de/cmmi_v1.2/browser.html#hs:null.
43. Young, H., Fang, T. and Hu, C. "A Successful Practice of Applying Software Tools to CMMI Process Improvement" *Journal of Software Engineering Studies*, 1(2): 78-95, December 2006.
44. Zurita-Rendón, H. "Arquitectura de la Herramienta Integral para MoProSoft". Universidad Nacional Autónoma de México. 2005 (in Spanish).

Security and Adaptability to Groupware Applications using a Set of SOA-based Services

Mario Anzures-García^{1,2}, Luz A. Sánchez-Gálvez¹, Miguel J. Hornos², and Patricia Paderewski-Rodríguez²

¹ Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla,
14 sur y avenida San Claudio. Ciudad Universitaria, San Manuel,
72570 Puebla, Mexico.

{anzures, luzsg}@correo.ugr.es

² Departamento de Lenguajes y Sistemas Informáticos, E.T.S.I. Informática y de
Telecomunicación, Universidad de Granada, C/ Periodista Saucedo Aranda, s/n,
18071 Granada, Spain.

{mhornos, patricia}@ugr.es

Abstract. Two fundamental aspects of groupware are security and adaptation. The former protects the information and the resources being shared, as well as allowing only those authorized users to make use of them. The latter allows us to adapt the application to the inherent dynamic of group work. This paper proposes a security model and an adaptability process, both of which use services based on SOA. As each service can be independently modified, both are adjusted according to the requirements of each group. Security model makes up of set services and focuses on controlling the user's access to shared resources (avoiding inconsistencies in groupware application) and the groupware application itself (preventing unauthorized users joining into or making use of it). The adaptability process focuses on the group organizational structure, using a set of services to adjust the Groupware application. In order to facilitate the inherent dynamic of the group organization structure is used an ontology, which to model the organizational structure, determining its behaviour through the concepts its relationship and the axioms defined.

1 Introduction

Groupware concentrates on support group work to achieve a common goal; therefore, it is necessary to facilitate group flexibility to respond to different collaborative scenarios that upsurge from the inherent dynamism of group work, as well as group interaction to allow their members to share resources and information, which is a main aspect to facilitate group work. Group flexibility depends on the adaptation of the group organization structure to the inherent dynamism of the group work; i. e. when users join or quit a session, when users change his/her role, when organization style is changed, or when a user plays several roles, etc. In sum, this dynamism is related to group size and organization, so that the application can be flexible enough, to adapt itself to the group requirements. When information and resources are shared,

the security is an important aspect to be considered. This happens in Groupware applications. Whereas in most systems, security is achieved through mechanisms such as: authentication, access control, data encryption, digital signature, etc. In the groupware field, special attention has been paid to authentication and access control mechanisms. The former is a mechanism that allows us to identify and verify the user's identity, trying to protect the system from unauthorized access. The latter is a mechanism that permits us to protect information according to security policies, by permitting access to shared resources only to authorized users.

This paper focuses on set services to provide a security model (which supports authentication and access control mechanisms to control interaction among users) and an adaptability process (that facilitates the adjustment to the group dynamic nature and to the changing needs of the same). Both avoid inconsistencies in the application on account of cooperative and competitive activities. In addition, SOA is the potential solution to the problems arisen from adaptation and reuse; therefore, it facilitates reuse and adaptation of each service and module here proposed. The paper is organized as follows: Section 2 describes the adaptability process. Section 3 explains the security model. Finally, Section 4 outlines the conclusions and future work.

2 Services-based Adaptability

Software adaptation [7] is based on the necessity of adjusting the system functionality in accordance with the new requirements that will appear in the future (changes in the environment, users' needs, different devices, etc.), in such a way that the system can continue working correctly. There are two forms of adaptability [6], [9]:

1. *Adaptive*, the adaptability is automatically performed and it is based on certain mechanisms previously defined by the designer and/or the developer.
2. *Adaptable*, the adaptability is carried out by the user's direct intervention in related to a set of constraints that avoid inconsistencies in the application.

The Groupware adaptability is focused, mainly, on the following aspects: access control [11]; concurrency control [4]; coupling of views [3]; and extensible architectures [8]. In the work presented in [1] these aspects have been addressed. In this paper, the adaptability focuses on the group organizational structure.

The adaptability controls in what way the components of the groupware application will be adapted when a change (or event) requiring modification takes place, so that the application functionality can be preserved. The adaptability of the group organizational structure comprises two phases: Pre-Adaptation and Adaptation.

2.1. Pre-Adaptation Module

This module (see Figure 1) provides a set of services which allow the Groupware application to determine whether it will carry out the application adaptability. Therefore, the Pre-Adaptation Module executes the following steps:

1. It monitors, at all times, (using *Detection Service*) events triggered in the execution environment (in this case, the *Collaborative Application Service*).

2. Each event detected by the *Detection Service* is compared with those contained in the *Adaptation Event Repository*.
 - 2.1. If a detected event is found into the *Adaptation Event Repository*, an adaptability process is required (go to step 3).
 - 2.2. Otherwise, an adaptability process is not required (go to step 1).
3. It determines the kind of adaptation to be carried out.
 - 3.1. Adaptable, go to step 4.
 - 3.2. Adaptive, take the following steps:
 - 3.2.1. A consensus must exist among all the group members (by means of the *Agreement Service*), to decide on (through the *Voting Tool Service*) whether an adaptation process should be performed or not.
 - 3.2.2. If the group decides to adapt the application, go to step 4.
 - 3.2.3. Otherwise, go to step 1.
4. Groupware application adaptability (go to Section 2.2).

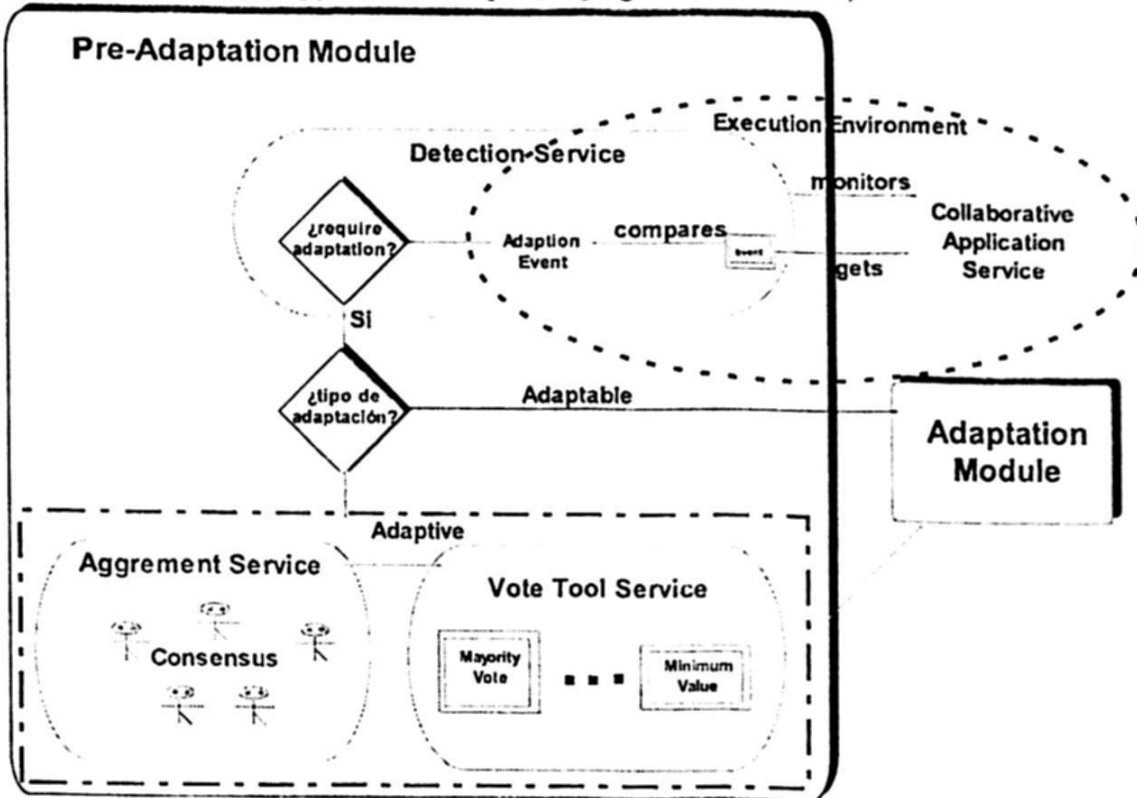


Fig. 1. Pre-Adaptation Module based on services.

2.1.1. Detection Service

This service (see Figure 1) monitors each event carried out at Groupware application through the *Notification Service*, which triggers itself whenever an event takes place. The *Detection Service* compares this event with those events registered at the *Adaptation Event Repository*, which contains only those events that involve an adaptability process. In addition, each event, in this repository, is associated to the kind of adaptability to be carried out, i.e., adaptive or adaptable process. The former

requires knowing the established organization structure in the group, namely that if it is a non-hierarchical organizational structure, the *Agreement Service* which allows users to reach a consensus using the *Vote tool Service*, will be executed. But, if it is a hierarchical organizational structure, an adaptable process is required.

2.1.2. Agreement Service

This service (see figure 1), is used by group members in an adaptive process whose organizational style is non-hierarchical, to decide whether the Groupware application adaptability should be performed or not. Therefore, the *Agreement Service* supports the consensus that all members have to reach.

2.1.3. Vote Tool Service

This service (see figure 1), provides a set of voting tools to be used by group members in order to reach an agreement. The *Vote Tool Service*, allows group members to choose from different kinds of agreements, which are based on a majority vote or on maximum/minimum value, etc. The group then will make a choice according to the established requirements to accomplish the given group task.

2.2. Adaptation Module

This module (see figure 2) performs the Groupware application adaptability; therefore, it executes a set of operations (which are stored in the *Operation Repository*) to determine that actions must be carried out in each service has been adapted. In order to avoid possible inconsistencies in the Groupware application, this phase is carried out by the *Adaptation Flow Service*.

2.2.1. Adaptation Flow Service

This service (see Figure 2) controls and manages the actions, which should be carried out in each service that will be adapted. These actions are determined by set of associated operations to an event. If some action can not be performed a reparation process is executed using the *Reparation Service*. The *Adaptation Flow Service* performs the following steps:

- a) It requires and gets to the *Adapted Components Service* the services list related to the adaptability process, as well as, the associated operations list to each service that show in the services list. The adaptability process starts with the first service and stop with the last service in the list, storing its name in the *Adaptation flow Repository* together with the operations that will allow us its adaptation. This adaptability process begins in the step "b)".
- b) It determines if the service meets the necessary pre-conditions (which establish the conditions that Groupware application must fulfil to be adapted)

- to be modified. If these pre-conditions are not satisfied the *Adaptation Flow Service* is stopped, a message that indicates the reason by which stopped is sent, and sometimes it is necessary perform a reparation process.
- c) It executes the adaptation operations associated with service to adapt it.
 - d) It decides if the service adapted meets the necessary post-conditions (which are the conditions that must be achieved later that the component has been modified). In case that the post-conditions are not satisfied, a reparation process is carry out, following the adaptation flow established by the adaptability process but in opposite direction. Thus, each component returns to their previous state and notifies users that adaptation can not take place.
 - e) The steps "b" to "d" are repeated until that all the services in the list got in the step a) have been adapted, thereby the adaptation process finish.

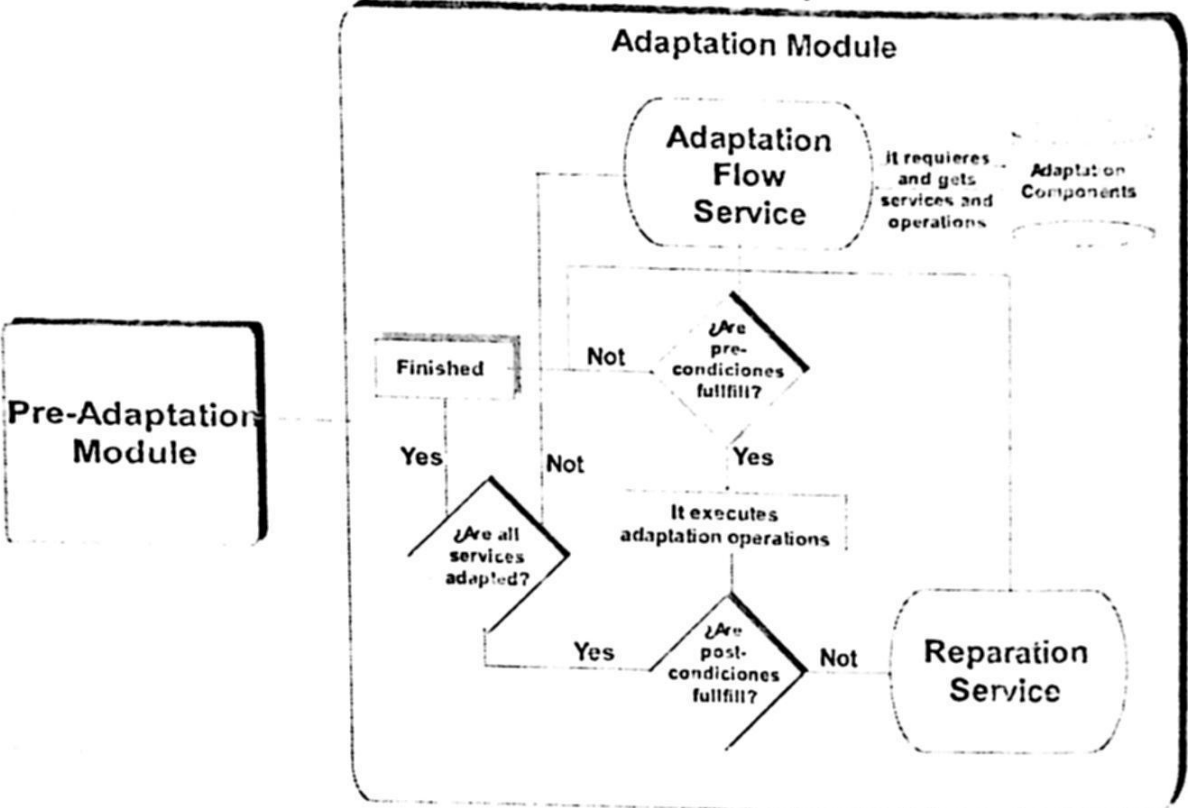


Fig. 2. Adaptation Module based on services.

2.1.2. Reparation Service

This service (see Figure 2) provides a mechanism in order to the application functionality can be preserved when the adaptability process can not be finished because not fulfill the post-conditions. The *Reparation Service* queries the *Adaptation Flow Repository* to get the last operation carried out in the last service adapted. In this way, each adapted service (starting with the last service and finishing with the first service of that repository) is returned to their previous state and notifies users that adaptability can not take place; preserving the appropriate application operability.

3 Services-based Security Model

In order to propose a service-based security model, on the one hand, the access control management and the authentication have been analyzed and modeled from the beginning of the Groupware application development. On the other hand, the main existent access control models have been studied; as a result we have detected their main problems, which are: they are not able to dynamically change permissions, they do not provide the global context of the collaborative application, and they do not specify neither restriction nor fine granularity control. The service-based security model is used to control how the shared resources are managed in the Groupware application in order to avoid inconsistencies. This model is based on RBAC model [10], the permissions are assigned to roles instead of users; in this way, when users modify their roles in the organization, the access control policy does not change.

The service-based model improves the existent access control models, for it considers: the static and dynamic aspects of the security, sophisticated and appropriate security policies, the general context related to collaborative activity, and the dynamic nature of the group, in such a way that permissions, roles and constraints are part of the model and these can be changed in runtime. In order to achieve these benefits, this model is based on four modules (see Figure 3): *Authorization*, *Session*, *Interaction Control*, and *Context*. Each of them abstracts a concern related to the establishing of sophisticated and appropriate access policies for group work, as well as, dynamic adaptation of its functionality in runtime. The changes can be:

- *Evolutionary*: Changes in the model which are dynamically updated in the collaborative application.
- *Adaptive*: Changes in the group organizational structure, which are predetermined in the access policy defined in the collaborative application.

3.1. Authorization Module

In the existent access control models, the authorization of users and the modification of permissions in runtime are core aspects, because access to resources in a session is controlled by the permissions or authorization that a user has at a given moment. This module (see figure 3) establishes the authorized roles to access a session, and defines how the users interact with shared resources and among themselves. It also allows the organizational structure to be dynamic. The *Authorization Module* facilitates an appropriate, dynamic and simple authorization of users in the session by four services: *Registration*, *Authentication*, *Stage*, and *Group Organizational Structure* (see Figure 3).

3.1.1. Registration Service

This service allows to the users to participate in a session, because the first thing that a user must do is to register her/himself on this. A user can join a group under own

petition or via an invitation. In the first case, the user must send a request via email indicating that she/he is keen to take part in the session. If the user meets the use conditions, the login and the password are sent to her/his email address. In the second case, the user receives in her/his email client an invitation with a login and a password to access the session. She/he will use the information received when she/he wants to join this session. When a user is registering in this, it is necessary that she/he fills in a registration form with her/his personal data and with useful information for carrying out the group work. The role that the user will play in this session is also stored and is provided in accordance with the valid roles in the *Group Organizational Structure Service* according to current stage (which is established in the *Stage Service*).

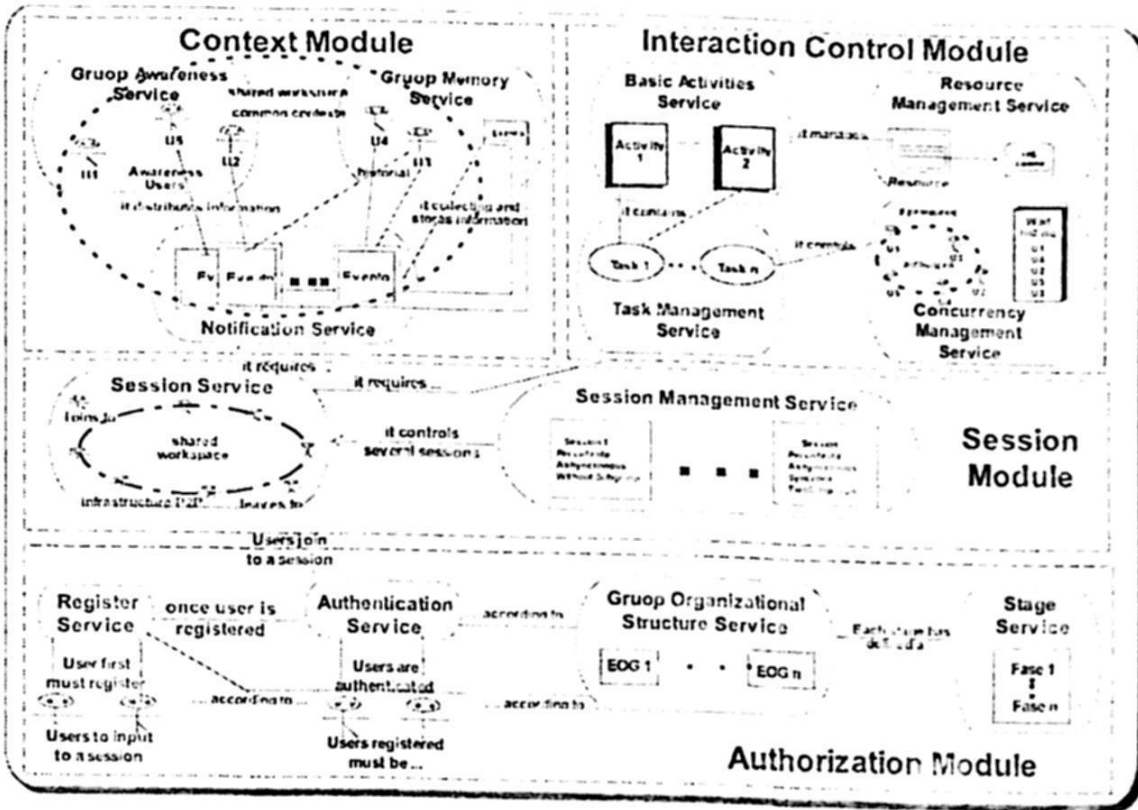


Fig. 3. Service-based Access Control Model

3.1.2. Authentication Service

This service controls the user access to the Groupware application, is used once the user is registered in the Groupware application. When the user inputs a login and a password, this service authenticates her/his access to the session, and stores these data, which are used by the system to corroborate that she/he is an authorized user. The authentication is dependent on the defined organizational style (*Group Organizational Structure Service*) and the currently executed stage (*Stage Service*).

3.1.3. Group Organizational Structure Service

This service presents an ontological model [2] that specifies the group organizational style taking into account all its static and dynamic aspects. Moreover, it allows controlling the access to the shared workspace and resources, and facilitates user authentication in this space. The group can change its organizational style in runtime. A style determines the roles that users can play and each role represents the set of access rights that users have on shared resources and the actions that they can perform. We have described the elements and the functional dependencies between them by means of this ontology, which allows specifying and controlling the changes on: the roles that a user can perform in a session, the access rights of each role, and the tasks to carry out in this. Moreover, in accordance with this model, we can create templates of policies that can be used to facilitate their reuse in runtime. For these reason, the security model facilitates the group work adaptability.

3.1.4. Stage Service

We consider long-term Groupware applications, where sharing information takes place at various stages. A stage [5] in a coordination model is defined as each of the collaboration moments. This service defines at least two stages: the collaboration and the configuration. The former, where the work group is carried out, can be in turn divided into several stages; for example, we distinguish four stages for it in the CMS: submission, assignment, review, and acceptance of papers. Each stage controls the roles that can participate in it, which facilitates the user authentication to the shared workspace. This service manages the collaboration moments, by establishing when a stage begins and ends, when a stage ends, it is possible to change the current organizational style (see section 3.1.3) to another more appropriate.

3.2. Session Module

This module supplies a shared workspace to carry out the interaction process in the Groupware applications, by means of two services (see Figure 3): *Session Management Service*, and *Session Service*.

3.2.1. Session Service

This service orchestrates the session, which includes its establishment, initiation, suspension, resumption and stopping, as well as the provision of information about the session state. A session denotes a set of geographically distributed individuals, who share a common interest to perform common tasks, so only a group can work in a session. This service provides a mechanism, which is supported by means of a peer-to-peer infrastructure, which allows connected users to work in a collaborative environment. In this way, it allows the users to join, to leave, to invite someone to,

and to exclude someone from a session. Once a session is established, this service manages and controls the session and the connections of users to it, storing information about each user and her/his work session; with this, it is possible to identify the users that are connected to each session. A user can participate in more than one session.

3.2.2. Session Management Service

This service supports the execution of several sessions at the same time. Here, each session is persistent, can be asynchronous or synchronous, and can be executed into another session. We consider session as persistent, which means they keep the session state until the next connection. This helps overcoming the well-known latecomer problem, and managing the adaptation process. Thanks to that, the session state can be retrieved from the server and it is possible to present the current state of the session to a new user. A Groupware application must be able to support both asynchronous and synchronous communication, so that they can be used when necessary. This service allows a group working in a session to make up a subgroup, which can start a new session. Both (group and subgroup) sessions can have different communication types, i.e. asynchronous or synchronous. This is important, because it allows the application to be adapted to the requirements of the group work.

3.3. Interaction Control Module

Once the group organization has been defined (i.e. user authorization to carry out the tasks and to use the shared resources) by means of the above mentioned module (see section 3.1.1), it is important to control the communication and coordination logic of the collaborative application to provide a good performance of the same at all times. This way, the mutually exclusive resource usage during the interactions among users must be guaranteed. This is achieved by means of this module, which contains four services, called: *Concurrency Management*, *Task Management*, *Basic Activities*, and *Shared Resource Management* (see Figure 3).

3.3.1. Concurrency Management Service

This service facilitates the manipulation of user permissions relevant to ensure exclusive access, which are granted in accordance with a default policy and a lock mechanism. The default policy for accessing the shared resources is "free for all", where the conflicts are resolved by a serialization of the access requests to shared resources on a first-come-first-served basis. This policy can be modified by the organizational style (established at the current stage). The access request to a resource is made to the *Shared Resources Repository*, which verifies whether the user has the required access rights to use that resource. In affirmative case, and if the resource is free, the user is allowed to use it to carry out the corresponding task. If the resource is

busy, the user requesting this resource is informed about the resource state and her/his request is put on the waiting list. If a user does not want to wait for the resource, she/he can remove her/his request from the waiting list at any moment. Once the resource is free (because the user who was using it finished their task or left the session), it is assigned to the next user on the list. This process is repeated until either all the users on the waiting list have used the resource or the session has finished. If the user does not have the rights to access to the resource, she/he receives a notification indicating that she/he can not use this resource.

3.3.2. Shared Resource Management Service

This service defines a flexible mechanism for storing the resources used which does not establish any restriction about the data logical model used in its description. The resources storing is performed through *XML* documents, so that the elements of the *Shared Resources Repository* are serialized in *XML* format. The *XML Schema* representation of each type of shared resource that the system uses needs to be stored in the meta-level repository. In this way, each invocation to the primitive contexts associated takes two parameters: one identifies the resource requested in the operation, and the other is a reference to its structural schema, so that each operation can be uniformly carried out regardless of the shared resource type on which it operates. Moreover, this service supplies a set of operations for managing resources in the repository. These operations allow us to insert, update, delete and query resources in the repository.

3.3.3. Basic Activity Service

This service defines a component that provides a set of activities to manage the resources stored in the *Shared Resources repository*. These activities are independent of the specific semantic associated with the different types of shared resources used, although their implementation should be carried out according to the data logical model used in the repository. This service uses the services of *Concurrency* and *Group Organizational Structure*.

3.3.4. Task Management Service

This service manages the tasks to be performed in the Groupware application. Each task is made up of a set of basic activities, which must be carried out to complete the corresponding task. The information about the tasks to be carried out is gotten from the *Group Organizational Structure Service*, which facilitates the tasks management. Due to the tasks are carried out in the shared workspace, this service uses the *Concurrency Management Service* that guarantees a coordinated management of the shared tasks used in the cooperative processes.

3.4. Context Module

The context is a very important element in the collaborative work, since the *Security Model* must take into account the current context of each user to establish the access permissions. Context information is necessary to manage the users and the resources access control, as well as users' authentication. This module (see Figure 3) contains three services, called: *Group Awareness*, *Group Memory*, and *Notification*, which provide a set of elements in relation to the collaborative context. Thanks to them, it is possible to control the state of each user, using the information stored in a session.

3.4.1. Group Awareness Service

This service provides to the users with the necessary information to support the group awareness. In this way, users are aware of other members presence in the session, as well as the actions that each of them has carried out and is carrying out, which facilitates user cooperation in collaborative tasks.

3.4.2. Notification Service

It provides users with operations for registering the interest in receiving collaborative events and for removing that interest, as well as for notifying the events produced during a session. For each user joining a session, an instance of the *Notification Service* is created. This instance is registered in the *Group Awareness* and/or *Group Memory Service* as an event consumer. Each cooperative event triggers the invocation of the corresponding notification operation, which delivers it to the *Group Awareness* and/or *Group Memory Service*, which in turn will propagate it to all registered consumers. These notifications are asynchronous, to avoid waiting and connections problems in the application. All events are distributed as strings, making the system suitable for any kind of event produced by any kind of Groupware application.

3.4.2. Group Memory Service

This service supplies a common context in the Groupware application which is called the group memory. This is a common space on which the collaborative activities of the group are carried out, and where the information about the shared resources is stored. This memory is created to provide understanding and reasoning about the collaborative process, and to do an exact tracking of this process. Therefore, this service collecting, storing and distributing information about the shared resources used and the activities undertaken with these resources, with the aim of supporting the dynamics of the group knowledge representation. In addition, the *Group Memory Service* provides persistence to the Groupware application, which gives the necessary information to perform several processes, such as the latecomer, adaptation or reparation ones.

4 Conclusions and Future Work

In this paper, it has been presented a set of SOA-based services for supporting the adaptability and the security in the Groupware applications. These are two quality attributes very important. The former is controlled and managed by five services, which allow us to adjust the organizational structure to the group necessities, preserving in all moment the appropriate application functionality. The latter is supported by *Security Model* using thirty services in order to protect the Groupware application, as well as the shared information and resources of unauthorized users. In this way, the services developed can be used in and/or modified to other applications to provide security and adaptability, thanks to that SOA provides reuse and adaptability. The future work will propose an ontology to facilitate the integration of these services in architectural model for develop Groupware applications.

References

1. Anzures-García, M., Hornos, M.J., Paderewski, P. Development of Extensible and Flexible Collaborative Applications using a Web Service-based Architecture, Springer-Verlag, LNCS Vol. 4401, 2007, pp. 66-80.
2. Anzures-García, M., Sánchez-Gálvez, L.A., Hornos, M.J., Paderewski-Rodríguez, P. Ontology-Based Modelling of Session Management Policies for Groupware Applications, Springer-Verlag, LNCS, Vol. 4739, 2007, pp. 57-64
3. Dewan, P., Choudhary, R. Coupling the User Interfaces of a Multiuser Program, ACM Transactions on Computer Human Interaction, Vol. 2-1, 1995, pp. 1-39.
4. Ellis, C.A., Gibbs, S.J. Concurrency Control in Groupware Systems, In Proceedings ACM SIGMOD, ACM Press, 1989.
5. Ellis, C., Wainer, J. A Conceptual Model of Groupware, In Proceedings of the ACM Conference on CSCW, 1994, pp. 79-88.
6. Fink, J., Kobsa, A., Nill, A. Adaptable and Adaptive Information for all Users, Including Disabled and Elderly People. The New Review of Hypermedia and Multimedia. Vol. 1-4, 1998, pp. 163-188.
7. Hiltunen, M.A., Schlichting, R.D. Adaptive Distributed and Fault-Tolerant Systems, Journal of Computer Systems and Engineering, Springer Verlag, Vol. 11-5, 1995.
8. Lee, J.H., Prakash, A., Jaeger, T., Wu, G. Supporting MultiUser, MultiApplet Workspaces in CBE, In Proceedings of 1996 the ACM Conference on CSCW, 1996, pp. 344-353.
9. Medina, N., García, L., Torres, J.J., Parets, J. Evolution in Adaptive Hypermedia Systems, In Proceedings of Conference on Principles of Software Evolution, 2002, pp. 34-38.
10. Sandhu, R. S., Coyne, E. J., Feinstein, H. L., Youman, C. E. Role-based Access Control Models, IEEE Computer, Vol. 29-2, 1996, pp. 38-47.
11. Shen, A.H., Dewan, A.P. Access Control in Collaborative Environments, In Proceedings of the ACM Conference on CSCW, 1992, pp. 51-58.

Using the CPAN Branch & Bound for the Solution of Travelling Salesman Problem

Mario Rossainz López¹, Manuel I. Capel Tuñón²

¹ Benemérita Universidad Autónoma de Puebla, Avenida San Claudio y 14 Sur,
San Manuel, Puebla, State of Puebla, 72000, México

rossainz@cs.buap.mx

<http://www.cs.buap.mx/~mrossainz>

² Departamento de Lenguajes y Sistemas Informáticos, ETS Ingeniería Informática,
Universidad de Granada, Periodista Daniel Saucedo Aranda s/n,

18071, Granada, Spain

manuelcapel@ugr.es

<http://lsi.ugr.es/~mcapel>

Abstract. This article presents the design of a High Level Parallel Composition or CPAN (according to its Spanish acronym) that implements a parallelization of the algorithmic design technique named Branch & Bound and uses it to solve the Travelling Salesman Problem (TSP), within a methodological infrastructure made up of an environment of Parallel Objects, an approach to Structured Parallel Programming and the Object-Oriented paradigm. A CPAN is defined as the composition of a set of parallel objects of three types: one object manager, the stages and the Collector objects. By following this idea, the Branch & Bound design technique implemented as an algorithmic parallel pattern of communication among processes and based on the model of the CPAN is shown. Thus, in this work, the CPAN Branch & Bound is added as a new pattern to the library of classes already proposed in [11], which was initially constituted by the CPANs Farm, Pipe and TreeDV that represent, respectively, the patterns of communication Farm, Pipeline and Binary Tree, the latter one implementing the design technique known as Divide and Conquer. As the programming environment used to derive the proposed CPANs, we use C++ and the POSIX standard for thread programming.

1 Introduction

At the moment the construction of concurrent and parallel systems has less conditioning than one decade ago, since there currently exist, within the realms of HPC or Grid computing, high performance parallel computation systems that are becoming more and more affordable, being therefore possible to obtain a great efficiency today in parallel computing without having to invest a huge amount of money in purchasing a state-of-the-art multiprocessor. Nevertheless, to obtain efficiency in parallel programs is not only a problem of acquiring processor speed; on the contrary, it is rather about

how to program efficient interaction/communication patterns among the processes [2],[6], which will allow us to achieve the maximum possible speed-up of a given parallel application. These patterns are aimed at encapsulating parallel code within programs, so that an inexperienced parallel applications programmer can produce efficient code by only programming the sequential parts of the applications [2],[4]. Parallel Programming based on the use of communication patterns is known as Structured Parallel Programming (SPP) [13], [14].

The present investigation centres its attention on the Methods of Structured Parallel Programming, proposing a new implementation, carried out with C++ and the POSIX Threads Library, as a CPAN [13],[14] of the algorithmic design technique known as Branch & Bound (BB). CPANs are Structured Parallel constructs based on the Object-Orientation paradigm useful to solve problems of high computational complexity by parallelizing their algorithms using a class of concurrent active objects. In this work the library of classes that we propose in [10] is complemented with the design and implementation of the CPAN Branch & Bound, which is intended to provide the programmer with an additional communication and interaction pattern among processes in parallel applications, which allows him to solve optimization problems, such as the Travelling Salesman Problem discussed here, which is an optimization problem with NP-Complete complexity.

2 Branch & Bound Method

Branch and bound (BB) is an algorithmic design technique that makes a partition of the solution space of a given optimization problem. The entire space is represented by the corresponding BB *expansion tree*, whose root is associated to the initially unsolved problem; the children at each node represent the subspaces obtained by *branching*, i.e. subdividing, the solution space represented by the parent node; and the leaves of the tree represent nodes that cannot be subdivided any further, thus providing a final value of the cost function associated to a possible solution of the problem. BB carries out a partial enumeration, i.e. a non-exhaustive search, over the nodes of the expansion tree until an optimal solution to the initial problem is found or the set of *live* nodes, i.e. those that still have the possibility of being branched, becomes exhausted. There are different possibilities to generate nodes and follow a route to a solution during the algorithm execution, known as the branching strategies of the BB algorithm, such as the ones given by the following search methods: *First in depth* (strategy LIFO), *First in width* (strategy FIFO) and *First best node*. The latter uses cost functions calculation to select the node that in principle seems to be more promising to explore, i.e. to further expand in order to find better solutions from it (strategy HFAP, using minimum cost or LC). In addition to these strategies, BB fixes bounds to the values of the suboptimal solution found at a certain point of the algorithm execution in order to prune those branches below a node that cannot lead to the optimal solution. A bound of the possible value of those reachable solutions is calculated in each node from the information contained in it. If the bound shows that any one of these solutions is necessarily worse than the best solution found up to that point, then there is no need for the algorithm to continue exploring on that branch and, therefore, prunes it off.

2.1 The Algorithm

Three stages are carried out in BB algorithms:

1. **Selection:** A node belonging to the set of live nodes is extracted. The selection directly depends on the strategy search which was decided for use in the algorithm.
2. **Branch:** the node selected in the previous step is subdivided in its children nodes by following a ramification scheme to form the expansion tree. Each child node receives from its father node enough information to enable it to search a suboptimal solution.
3. **Bound:** Some of the nodes created in the previous stage are deleted, i.e. those whose partial cost, which is given by the cost function associated to this BB algorithm instance, is greater than the best minimum bound calculated up to that point.

The contribution of this algorithm is that it offers a way to perform the greatest possible reduction of the space search and, therefore, obtains a decrease in the exploration complexity of the expansion tree which contains the optimal solution being searched. The nodes that have not yet been pruned are included in the live node list, and thereby, the selection process begins again until the algorithm finalizes. The general structure of the algorithms that implement the BB technique is based on three main modules:

1. The module that contains the scheme of general operation of the technique.
2. The module that represents the structure of data where the generated nodes are stored.
3. The module that describes the structures of data that conform the nodes.

The first module is the only one that remains without modification, independently of the problem to solve with the BB algorithm and it is valid for all the algorithms that follow the technique. The pseudo-code implementing it is as follows:

CLASS CONCRETE EsquemaBB

```
{
  Estructura e;
  Nodo n;
  Nodo[] hijos;
  int numhijos,i,j;
  PUBLIC Nodo B&B()
  {
    e= Estructura CREATE();
    n= nodoInicial();
    e.inserta(n,n.h());
    WHILE (!E.esVacia())
    {
      n=e.extrae();
      numhijos=n.expandir(VAR hijos);
      eliminar(n);
      n.poner_cota_sup(numhijos,hijos);
      FOR i=(0,numhijos)
      {
        IF (n.aceptable(hijos[i]))
        {
```

```

    IF (n.esSolucion(hijos[i]))
    {
        FOR j=(0,numhijos)
            IF (i!=j)
                DELETE hijos[j];
        e.clear();
        RETURN hijos[i];
    }
    ELSE
        e.inserta(hijos[i],
                  n.h(hijos[i]));
}
ELSE
    DELETE hijos[i];
}
}
}

```

The definition of the abstract data type that represents the structure *e* where the nodes are stored corresponds to the ADT HEAP, because the strategy used to search a node containing a solution is that of selecting the minimum cost (LC) one among the contained in the HEAP. The definition of the class that represents the type *Node* used in the previous pseudo-code is composed of the following functions:

- *expandir()*: It is the function that creates the children nodes out of a given node and returns the number of children to where the function is called. This function is the one that implements the process of node ramification of the algorithm
- *aceptable()*: Function that carries out the pruning of unpromising nodes and, when it obtains a live node, decides whether to continue exploring or to reject it.
- *esSolucion()*: It is a function that decides whether its parameter node is a leaf of the tree, that is to say, a possible solution to the original problem.
- *h()*: This function in its two versions, i.e. it can be overloaded, is the one that implements the cost function for the search strategy LC and its value is used as a priority position value when storing the nodes in the HEAP structure.
- *poner_cota_sup()*: It allows for the upper bound of the problem to be established. The function that carries out node pruning uses this datum to prune those nodes whose cost value is greater than the currently obtained bound of the optimal solution.

The CPANS can provide the parallel algorithms necessary to solve problems, such as the one of the TSP, using the BB technique. When solving this kind of problems, in addition to the solution, reasonable run times of the whole computation can be obtained with the CPAN model. Since the complexity in NP-complete problems as in the TSP one is intrinsic, parallelization is the only way to obtain a solution in the practice. Each object node of the expansion tree therefore needs to be independent that is to say, it must contain all the necessary information to be an active object, (i.e. to have the capability of execution in itself) which makes it possible for the processes of branch and bound to perform the reconstruction of the solution found up to that moment.

3 The Travelling Salesman Problem

The Travelling Salesman Problem could be represented by a directed graph consisting of a set of vertices (cities) and labelled arcs (distances between cities). One optimized solution of the TSP is a path in which all the vertices have been visited exactly once with minimal cost [3]. Formally, the problem can be enunciated as follows: given a connected and weighted graph g and given one of its vertices v_0 as the initial one, we must find the Hamiltonian cycle of minimum cost that begins and finishes in v_0 [8].

3.1 Solution of the TSP with the BB technique

The problem is solved by a BB algorithm, which dynamically builds a search tree, whose root is the initial problem and its answer nodes are complete tours [3] around all the cities represented by the nodes of the graph. Numerous strategies of Branch & Bound exist that solve the problem of the TSP. The first of the three strategies described in [15] is used in this work where the following important elements are defined, as well:

- **$LC(P)$ - Cost of a node P :** Is the distance of a complete graph's tour after P gets included in it, if P is a solution node, or, otherwise, is the cost of a solution of minimal cost in the sub-tree whose root is P .
- **$cota$ - Upper bound:** The length of the shortest complete tour found up until that moment. This value must be a global variable in the program and is used to prune the search tree of nodes.
- **$h(P)$ - lower bound** of the cost of a solution for a problem or sub-problem P . If P is a solution node $LC(P) = h(P)$.
- **$s[]$ - vector solution:** It is a vector that indicates the order in which the vertices must be visited to reach the optimal solution. Each element of the vector contains a number between 1 and N , being N the number of vertices of the graph that defines the problem. The vector cannot contain repeated elements.
- **$M[[]]$ - Matrix of adjacency:** It is the representation of the graph where the vertices are the indices of the matrix and the contents of the matrix elements given are the arcs between two vertices. The matrix of adjacency is not necessarily symmetrical, although, it is so with respect to its nonnegative elements.

Computing the lower bound of a node of the search tree is carried out by obtaining the *reduced cost matrix* for each node [3]. A row (or column) of a matrix is reduced if it contains an element zero at least, and the rest of the elements are nonnegative. We say that a matrix is reduced if and only if all its rows and columns are reduced [8]. With respect to the interpretation of the cost of a node, this is obtained by adding the amounts t_i in which the rows or columns of the matrix of adjacency of the graph are reduced when the process of obtaining the reduced matrix is carried out. This amount which is removed when reducing a matrix is a lower bound of the total cost of any possible tour traversing the graph nodes. This is exactly what it is used as, the cost function LC to prune nodes p of the expansion tree. Therefore, a *reduced matrix* and an *accumulated cost* are associated to each node and if we suppose that $M[[]]$ is the

reduced matrix associated to the node p and p' is a children node of p which is obtained by including the arc $\{i,j\}$ in the partially built tour $s[]$, then:

- If p' is a leaf node of the tree, i.e. a possible solution, its cost is the accumulated cost of p plus $M[i,j] + M[j,1]$. The latter term is the one that completes the tour. The amount obtained is the cost of such a tour.
- If p' is not a leaf, its reduced matrix M obtained from matrix M' assigned to this node p' will have as cost the cost of p added to the cost of the reduction of M' added to the value of $M[i,j]$.

4 Definition of CPAN

The concept of CPAN has mostly been carried out within the PhD thesis research work referenced in [11]. The basic idea is implementing different types of parallel patterns of communication between the processes of an application and implementing distributed/parallel algorithms as classes, by following the Object-Orientation paradigm. The execution of a method of the objects that constitute a CPAN can be carried out through a message sent to an instantiated class, which acknowledges it as a service petition.

A CPAN comes from the composition of a set of parallel objects [12] of three types (see Fig 1):

An object manager representing the CPAN itself and makes of it an encapsulated abstraction that hides its internal structure. The manager controls the references of a set of objects (a denominated object Collector and several denominated Stage objects) that represent the components of the CPAN and whose execution is carried out in parallel and should be coordinated by the manager itself.

The Stage objects are objects of specific purpose responsible for encapsulating a client-server type interface between the manager and the object slaves (objects that are not actively participative in the composition of the CPAN, but rather, are considered external entities that contain the sequential algorithm that constitutes the solution of a given problem) as well as providing the necessary connection among them to implement the communication semantic pattern that seeks to be defined. In other words, each stage should act in parallel as a node of the graph that represents the communication pattern and should be capable of executing its methods as an active object. A stage can be directly connected to the manager and/or to another component stage depending on the pattern peculiar to the CPAN being implemented.

And an object Collector which is an object in charge of storing in parallel the results that it receives from the stage objects that are connected to it. That is to say, during the service of a petition, the control flow within the stages of a CPAN depends on the implemented communication pattern. When the composition concludes its execution, the result does not return to the manager directly, but rather to an instance of the class Collector which takes charge of storing these results and of sending them to the manager, which will then send them to the exterior, as soon as they arrive to it, without the need of waiting for all the results to be obtained.

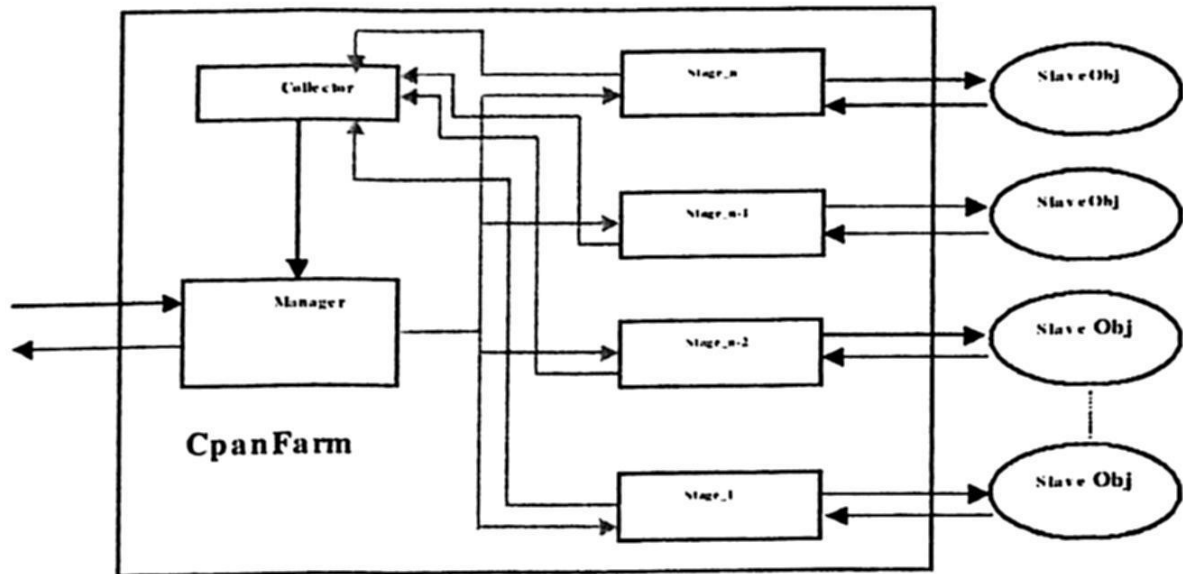


Fig 2. The Cpan of a Farm communication pattern [10], [12]

5 Parallelization of Branch & Bound Technique

The ramification is separated from the bounding of nodes on the expansion tree when the BB algorithm executes. These two structures were implemented using the *Cpan Farm* of the library proposed in [10], so that the ramification and the distribution of work to the processes were carried out by using the *scheme of the Farm*. As Fig 3 shows, the expansion tree, for a given instance of the BB algorithm, is obtained by iteratively subdividing the stage objects according to the farm pattern until a stage representing a leaf-node of the expansion tree is found, i.e., one stage in charge of solving a sub-problem that cannot be additionally subdivided.

On the other hand, the pruning is carried out implicitly within a *farm* construction by using a *scheme totally connected* between all the processes. It can communicate a sub-optimal bound found in a process to all the processes that are branching to avoid ramifications of useless sub-problems, i.e., those that do not lead to improving the best upper bound obtained up to that moment.

The *Cpan Branch & Bound* is composed of a set of *Cpans Farm* that represent worker processes and a controller, therefore, forming a new type of Farm, the *Farm Branch & Bound* or *FarmBB* that will be included in the library of CPANS. The *Cpan Farm* workers are executed in parallel forming the expansion tree of nodes given by this technique. The process controller of the initial *Cpan Farm* represents the root of the expansion tree that is in charge of distributing the work and of controlling the progress of the global calculation given to the collector of the *FarmBB* which sends the result to the process controller of the *Cpan FarmBB*, which then shows it to the user. See Fig 3.

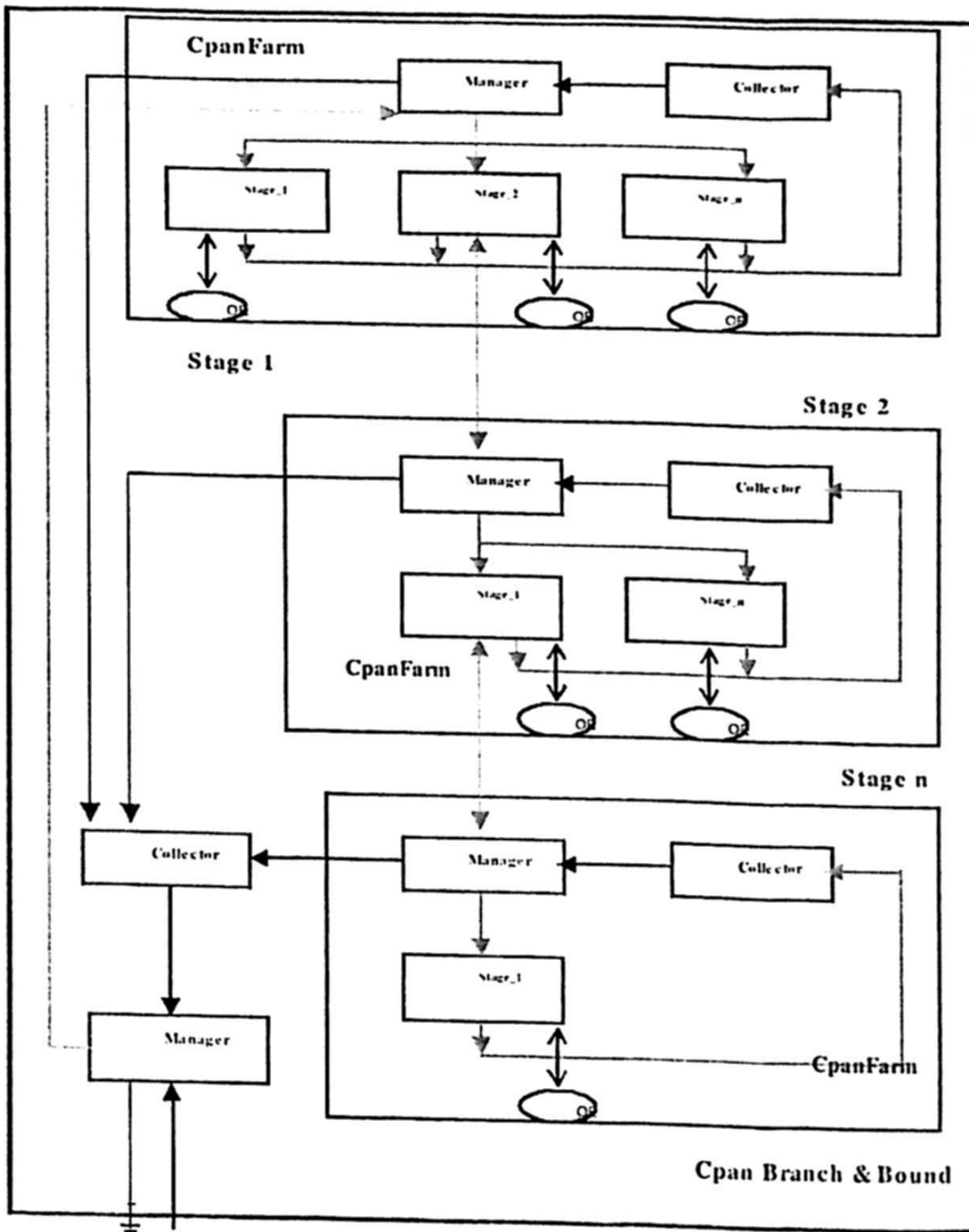


Fig 3. The Cpan Branch & Bound [11]

6 Results and Analysis of Speedup

The search strategy which was used in the implementation and test of the distributed TSP was: The first best search strategy that uses the calculation of cost functions for each live node to select the node that, in principle, seems the most promising to analyze (strategy HEAP, using minimum cost or LC).

The analysis of Speedup of the CPANS B&B that appears in table 1 and Fig 4 was carried out in a Parallel System Origin 2000 Silicon Graphics (of 64 processors) available in the European Center for Parallelism of Barcelona CEPBA.

Table 1. Execution of Parallel CpanB&B in 2, 4, 8, 16 and 32 processors with N=50 cities

	CPUSEQ	CPU2	CPU4	CPU8	CPU16	CPU32
Run time	35.42 Seg.	21.88 Seg.	14.21 Seg.	11.34 Seg.	10.27 Seg.	9.10 Seg.
Time CPU	27.10 Seg.	23.25 Seg.	21.17 Seg.	19.19 Seg.	22.69 Seg.	22.18 Seg.
CPI	1.321	0.952	0.943	0.928	0.924	0.914
Speed Up	1.00	1.62	2.49	3.12	3.45	3.89
Amdalh	1.00	1.68	2.55	3.43	4.16	4.64

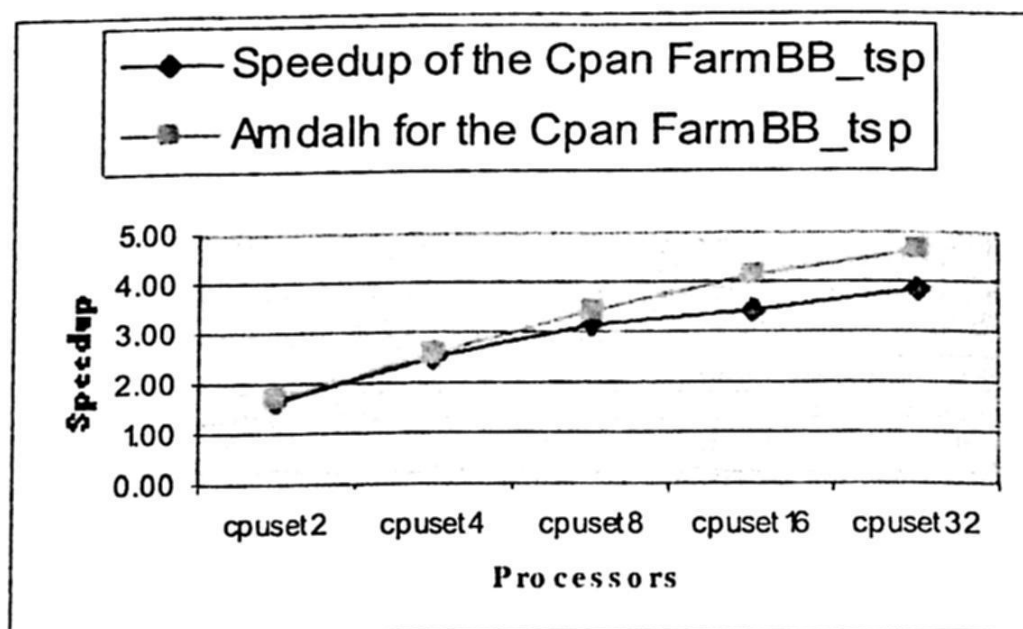


Fig 4. Speed Up of Parallel Cpan B&B

7 Conclusions

1. The technique of Branch & Bound as a High Level Parallel Composition or CPAN has been implemented.
2. The utility of the library of CPANS proposed in [10] which serves to make compositions of CPANS and to define new CPANS models as in the *Cpan Branch & Bound* has been demonstrated.
3. With the model of the *Cpan Branch & Bound* we have been able to offer an optimal solution of a TSP NP-Complete problem.
4. The CPANS Pipe, Farm, TreeDV and Farm-Branch-&-Bound constitute the library of classes of the Cpan.

References

1. Brassard G.; Bratley P. 1998. "Fundamentos de Algoritmia". Prentice Hall. España. ISBN: 84-89660-00-X.
2. Brinch Hansen; "Model Programs for Computational Science: A programming methodology for multicomputers". *Concurrency: Practice and Experience*, Volume 5, Number 5, 407-423, 1993.
3. Capel M., Palma A. 1992. "A Programming tool for Distributed Implementation of Branch-and-Bound Algorithms". *Parallel Computing and Transputer Applications*. IOS Press/CIMNE. Barcelona.
4. Capel M., Troya J. M. "An Object-Based Tool and Methodological Approach for Distributed Programming". *Software Concepts and Tools*, 15, pp. 177-195. 1994.
5. Ceballos F.J. 2003. "Programación Orientada a Objetos con C++". Editorial RA-MA. España. ISBN: 84-7897-570-5.
6. Darlington et al. "Parallel Programming Using Skeleton Functions". *Proceedings PARLE'93*, Munich (D), 1993.
7. Goodman S.E.; Hedetniemi S.T. 1997. "Introduction to the Design and Analysis of Algorithms". Mc Graw Hill Book Company. United States of America. ISBN:0-07-023753-0.
8. Guerrequeta G.R.; Vallecillo M.A. "Técnicas de Diseño de Algoritmos". *Manuales*. Universidad de Málaga.
9. Troya J. M., Capel M. "An Object Based Tool for Distributed Programming on Transputer Systems".
10. Rossainz M., Capel M. "A Parallel Programming Methodology based on High Level Parallel Compositions (CPANs)". *Proceedings of XIV International Conference on Electronics, Communications, and Computers, CONIELECOMP*. ISBN 0-7695-2074-X.
11. Rossainz M. "Una Metodología de Programación Basada en Composiciones Paralelas de Alto Nivel (CPANs)". Universidad de Granada, PhD dissertation, 02/25/2005.
12. Rossainz M., Capel M. "An Approach to Structured Parallel Programming Based on a Composition of Parallel Objects". *Congreso Español de Informática CEDI-2005. XVI Jornadas de Paralelismo*. Granada, Spain 2005. Editorial Thomson. ISBN: 84-9732-430-7.
13. Corradi A.; Zambonelli F. 1995. "Experiences toward an Object-Oriented Approach to Structured Parallel Programming". DEIS technical report no. DEIS-LIA-95-007.
14. Danelutto, M.; Orlando, S; et al. "Parallel Programming Models Based on Restricted Computation Structure Approach". *Technical Report-Dpt. Informatica*. Università de Pisa.
15. Horowitz, Sahni. 1978. "Fundamentals of Computer Algorithms". Ed. Computer Sc. Press.

Fast Automatic Retinal Blood Vessel Segmentation and Vascular Landmarks Extraction Method for Biometric Applications

Fabiola M. Villalobos-Castakli¹, Edgardo M. Felipe-Riverón²

^{1,2} Center for Computing Research, National Polytechnic Institute,
Av. Juan de Dios Batiz and Miguel Othon de Mendizabal, Mexico,
D.F., P. O. 07738. Mexico, Phone: 5729 6000/56515.
fvillalobosb07@sagitario.cic.ipn.mx,
edgardo@cic.ipn.mx

Abstract. The retina has many desirable characteristics as the basis of authentication. Retinal blood vessel patterns are known to be very distinctive, even between identical twins. The blood vessel structure is very stable over time, well shielded from outside environmental impacts, and believed to be difficult to spoof. Retinal Identification seeks to identify a person by comparing images of the blood vessels in the back of the eye, the retinal vasculature. This method takes advantage of the fact that of all human physiological features, the retinal image is the best identifying characteristic. This article describes a fast, efficient and automatic algorithm for segmenting retinal blood vessels and for extracting vascular landmarks from these vessels as a unique representation used for biometrical applications. The proposed segmentation method is based on the second local entropy and on the gray-level co-occurrence matrix (GLCM). The algorithm is designed to have flexibility in the definition of the blood vessel contours. Using information from the GLCM, a statistic feature is calculated to act as a threshold value. After the segmentation stage, a morphological thinning process is applied and the landmarks are detected and their attributes are extracted. Finally the "eye print" representation is constructed using this salient features. The results obtained show the effectiveness and accuracy of the proposed method to detect and extract information from a retinal fundus images. The elapsed time for the proposed segmentation method is 3.2 seconds.

Keywords: Biometric identification; Retinal recognition; Retinal vessel tree; Bifurcation and ending points; Image segmentation; Co-occurrence matrix; Entropy thresholding.

1 Introduction

The terms *Biometric* and *Biometry* have been used since early in the 20th century referring to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences [1]. For a layman, it

could be said that Biometry is the science of measuring physical and/or behavioral characteristics that are unique to each individual and could be used to verify that an individual is who he or she claims to be. Since these characteristics are unique to each individual, biometrics are believed to effectively combat theft and fraud in a wide variety of industries and applications. Notably, the recent advances of information technology and the increasing requirement for security have led to a rapid development of intelligent personal identification system based on biometrics [12].

The retina has many desirable characteristics as the basis of authentication. Retinal blood vessel patterns are known to be very distinctive [2], even between identical twins [3]. The blood vessel structure is very stable over time, well shielded from outside environmental impacts, and believed to be difficult to spoof [6]. Retina-based identification has long been perceived as a robust biometric solution but very few practical applications or commercially viable products have been demonstrated. However, it suffered from a human interface perceived as intrusive and unfriendly. The genesis of the retinal identification technology lies in the medical imaging field. Retinal imaging devices and automated diagnostic tools were developed for a range of retinal disease states. This led to an understanding of imaging the retinal blood vessel network and an interest in developing automated tools for its analysis [9].

2 Retinal blood vessels as a biometric

Retinal identification (RI) is an automatic method that provides true identification of persons by acquiring an internal body image, the retina of a willing person, who must cooperate in a way that would be difficult to counterfeit [6]. Awareness of the uniqueness of the retinal vascular pattern dates back to 1935 when two ophthalmologists, Drs. Carleton Simon and Isidore Goldstein, while studying eye disease, noted that every eye has its own totally unique pattern of blood vessels. They subsequently published a paper on the use of retinal photographs for identifying people based on blood vessel patterns [2]. Later in the 1950s, their conclusions were supported by Dr. Paul Tower in the course of his study of identical twins. He noted that, of any two persons, identical twins would be the most likely to have similar retinal vascular patterns. However, Tower's study showed that of all the factors compared between twins, retinal vascular patterns showed the least similarities [3].

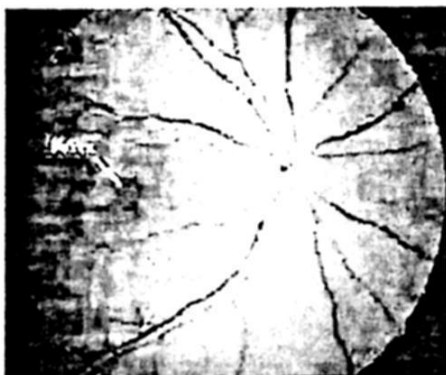


Fig. 1. Retinal blood vessels appearance.

2.1 Retinal Blood Vessels Characteristics

It is the blood vessel pattern in the retina (Figure 1) that forms the foundation for the science and technology of the RI [6]. This method takes advantage of the fact that of all human physiological features, the retinal image is the best identifying characteristic [5]. Because of the complex structure of the vessels that supply the retina with blood, each person's retina and also each person's eye is unique.

The blood vessels of the vascular network of the retina have the following characteristics:

a) Uniqueness. It is unique in:

- The number of major blood vessels that are located in the area of the optic nerve [5], [9], [25]. The biometric systems that detect a vascular pattern of an individual's retina to identify it use the vascular structure outside of the optic disc because it was thought that only this area of the retina contained enough information to distinguish one individual from another [7].
- The relative angle of these major blood vessels as they emerge from the optic disc. The central retinal artery and vein can be seen to bifurcate rapidly at the optic disc [9].
- The branching characteristics of the blood vessels. Among all features, bifurcation points are the most reliable and abundant features in the fundus images. It has been suggested that non-equilibrium Laplacian process could be involved in retinal angiogenesis and that fluctuations in the distribution of embryonic cell-free spaces provides the randomness needed for fractal behavior and for the uniqueness of each individual's retinal vascular pattern [9].
- The position and size of the optic disc. The optic disc, seen as the bright spots in Figure 1, is the point where the optic nerve breaks out into the retina. This disc is approximately $15.5^\circ \pm 1.1^\circ$ nasals and $1.5^\circ \pm 0.9^\circ$ superior to the fovea [32].
- The pigments or coloring patterns of the retina and of the retinal vasculature.
- The infinite variability that exists with respect to certain anatomical landmarks of the retinal vasculature [4]. Its detailed final structure is mostly stochastic and thus its uniqueness stands of reasons [8].

Also, the retinal image has some other characteristics that place it as the best biometric identification option.

b) Permanence. The retinal image does not change significantly with time. Other than cases of significant trauma, pathology, or biochemical interference, spontaneous adult ocular vasculogenesis and angiogenesis usually do not occur [8]. Age or disease may change the characteristics of the eye blood vessels, but not their position in the retina [11].

c) Reliability. It is impossible to counterfeit the retinal image. The fine, multi-surface structure of the ocular vessels makes them hard to reproduce as a physical artifact [8]. Fraud-proof, that is, it is virtually impossible to replicate the image produced by a human retina for unauthorized access to computer networks, medical records or physical facilities [10]. Moreover, imaging the retinal vasculature may be done to the eyes of living persons. Therefore, by accurately recording and analyzing the

configuration of the blood vessels, the subject can be positively being identified. Since the configuration of retinal blood vessels of an eye can not be falsified or altered, it offers an incredibly, accurate, inalterability, and unchanging characteristic of the subject [4].

d) Universality. A holangiote eye is an eye having vasculature on the ocular fundus, with the vasculature entering the eye, primarily, through the optic nerve head. Humans, as well as virtually all domestic animal species and many game animal species, including deer and elk, have holangiote eyes. Based on this, it is possible to identify an individual using the image that is acquired of an eye of the animal (human or non-human) and the retinal vasculature is extracted from that image [4]. In humans, the retinal vasculature disappears within seconds of the cessation of life, thereby insuring that the captured image was obtained from a living subject.

e) Safety. The eye shares the same stable environment as the brain and among physical features unique to individuals; none is more stable than the retinal vascular pattern. Because of its internal location, the retina is protected from variations caused by exposure to the external environment (as in the case of fingerprints, palm prints, etc.).

f) Representative. By targeting common structures such as the optic disc and the retinal vascular branches, a consistent source of readily identifiable, yet contrasting structures are available for digital imaging and processing.

g) Certainty. Unlike passwords, a retina cannot be forgotten; unlike plastic cards, a retina cannot be lost or loaned by someone else.

The majority of the disadvantages that owns the biometric identification based on the vascular network of the retina are inherent to the user interface.

As mention before, retinal based recognition for personal identification has desirable properties such as uniqueness, stability, permanence, etc. However, research on retinal vessel structure for biometric applications, has not revealed its full potential. Retinal vessel structure has extraordinary structures and provides many interlacing characteristics, which are unique for each person, so it will be one of the most reliable and accurate biometric [12].

3 The proposed method

As depicted in Figure 2 the proposed method is composed of 3 main processing stages: 1) a pre-processing step, 2) a main process step, and 3) a post-processing step. The pre-processing step consists of the following 3 stages: a) green-color band selection, b) mask generation, and c) image enhancement for vessel network detection. The main process consists of 4 stages: a) co-occurrence matrix computation, b) vessel segmentation by the second entropy thresholding technique, c) morphological thinning, and d) landmarks detection. Finally, the post-processing step contains 2 stages: a) pruning and b) landmark attributes estimation.

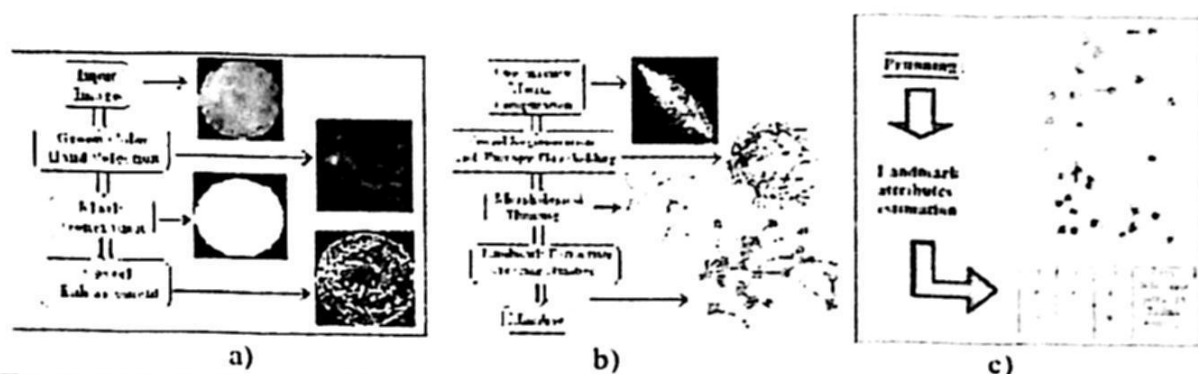


Fig. 2 a) Block diagram of the proposed method; a) Pre-processing step, b) Main process stage, and c) post-processing step.

3.1 Green color band selection

A gray-level image is produced by extracting the green layer of the original RGB image. The green component of color image gives the blood vessels on a highly contrasted background (dark blood vessels on a bright background). Hence, the green channel of image is employed in the retinal vasculature detection [12].

3.2 Mask generation

Mask generation aims at labeling pixels belonging to the fundus Region of Interest (ROI) in the entire image [31]. Pixels outside that ROI are those belonging to the dark surrounding region in the image.

3.3 Image enhancement

Denoising, correction of the brightness and contrast enhancement are applied in this pre-processing step [13]. Therefore, the proposed enhancement method detects vessels using the knowledge of their known gray level profile and the concept of the matched filter detection, which is used to detect piecewise linear segments of blood vessels in retinal images.

3.3.1 Bell-Shaped Gaussian Matcher Filter (BSGMF)

It can be noted that the retinal vessels can be represented by piecewise linear segments with Gaussian-shaped cross sections. A matched filter is constructed for the detection of the vessel edge segments searching in all possible directions. A Bell-Shaped Gaussian matched filter (BSGMF) was developed to cover all 12 orientations where designed kernel was given by Eq. 1 [16].

$$K(x, y) = \pm \exp\left(-\frac{x^2 + y^2}{2s^2}\right) \quad (\text{Eq. 1})$$

With the tail truncated at $x^2 + y^2 = 3s^2$. The application of this method enhances individual vessels segments in the image. A proper thresholding scheme must be used to distinguish between the enhanced vessel segments and the background.

3.4 Computation of the co-occurrence matrix

A co-occurrence matrix of an image is an $L \times L$ square matrix, denoted by $W = [t_{ij}]_{L \times L}$ whose elements are specified by the numbers of transitions between all pairs of gray-levels in $G = \{0, 1, \dots, L-1\}$ in a particular way [33], [34]. Each entry in the matrix t_{ij} gives the number of times the pixel gray-level j follows the gray-level i in some pattern [17], [18].

3.4.1 Quadrants of the co-occurrence matrix

Let t be a value used to threshold a gray-level image. It partitions a co-occurrence matrix into four quadrants, namely, A, B, C and D. We assume that pixels with levels above the threshold are assigned to the foreground (corresponding to vessels), and those equal to or below the threshold are assigned to the background. Then, quadrants A and C correspond to local transitions within foreground and background, respectively, whereas quadrants B and D are joint quadrants which represent joint transitions across boundaries between background and foreground. The probabilities of the gray-level transition within each particular quadrant can be by the so called 'cell probabilities' (Eq. 2):

$$P'_{a|a} = \frac{P_{aa}}{P'_a} \quad P'_{a|b} = \frac{P_{ab}}{P'_b} \quad P'_{b|c} = \frac{P_{bc}}{P'_c} \quad P'_{b|d} = \frac{P_{bd}}{P'_d} \quad (\text{Eq. 2})$$

3.5 Blood vessel segmentation

The objective of retinal vessel segmentation is to decide which part of the image belongs to the foreground (which is of our interest for extracting features for recognition and identification), and which part belongs to the background (which is the noisy area around the boundary of the image) [14]. Reliable vessel extraction is a pre-requisite for subsequent retinal image analysis and processing because vessels are the predominant and most stable structures appearing in image [12]. Accurate segmentation of retinal images influences directly the performance of minutiae extraction. If more background areas are included in the segmented retinal image, more false features are introduced; if some parts of the foreground are excluded, useful feature points may be missed [14]. The first step of 2-D segmentation is to build the 2-D histogram, which can be commonly established by using the gray-level co-occurrence matrix [19].

3.5.1 Gray-level co-occurrence matrix used for relative entropic thresholding

Relative entropy has been used to measure the information distance between two information sources. The smaller the relative entropy is, the closer the two sources are in terms of their probability distributions. The transition probabilities defined by the co-occurrence matrix contain the spatial information that reflects homogeneity of local gray-level transitions in quadrants A and C, and joint gray-level transitions across boundaries in joint quadrants B and D [20], [21].

Let the second-order relative entropy of the gray-level transition probabilities $\{p_{ij}\}_{i=0, j=0}^{L-1, L-1}$ and $\{h'_{ij}\}_{i=0, j=0}^{L-1, L-1}$ be defined by:

$$J(\{p_{ij}\}; \{h'_{ij}\}) = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{ij} \log \frac{p_{ij}}{h'_{ij}} \quad (\text{Eq. 3})$$

Where p_{ij} are the transition probabilities from gray level i to gray level j of the original image and h'_{ij} is the transition probability generated by the thresholded binary image in response to p_{ij} . Using Eq. 3 as a thresholding criterion to minimize $J(\{p_{ij}\}; \{h'_{ij}\})$ over t generally renders the thresholded binary image that best matches the original image.

3.5.2 Local Relative Entropy (LRE) thresholding

If we define $P'_{i|ic} = \frac{P_{ij}}{(P'_i + P'_c)}$, and normalize the probabilities in the local quadrants A and C, then we get:

$$J_{LRE}(\{P'_{i|ic}\}; h'_{ij}) = H_{AB+CD}(t) - \sum_{i,j \in AB \cup CD} P_{i|ic} \log h'_{ij} \quad (\text{Eq. 4})$$

Where

$$H_{AB+CD}(t) = - \sum_{i,j \in AB \cup CD} P_{i|ic} \log P_{i|ic} \quad (\text{Eq. 5})$$

is the entropy of local quadrants A and C in the co-occurrence matrix W . The second term in Eq. 5 can be further reduced to:

$$\sum_{i,j \in AB \cup CD} P_{i|ic} \log h'_{ij} = \frac{P'_i}{P'_i + P'_c} \log \left(\frac{q'_i}{P'_i + P'_c} \right) + \frac{P'_c}{P'_i + P'_c} \log \left(\frac{q'_c}{P'_i + P'_c} \right) \quad (\text{Eq. 6})$$

Substituting Eq. 6 into Eq. 4 results in:

$$J_{LRE}(\{P'_{i|ic}\}; h'_{ij}) = -H_{AB+CD}(t) \left[\frac{P'_i}{P'_i + P'_c} \log \left(\frac{q'_i}{P'_i + P'_c} \right) + \frac{P'_c}{P'_i + P'_c} \log \left(\frac{q'_c}{P'_i + P'_c} \right) \right] \quad (\text{Eq. 7})$$

The LRE thresholding method aims to find a threshold value t_{LRE} that minimizes $J_{LRE}(\{P'_{i|ic}\}; h'_{ij})$, that is:

$$t_{LRE} = \arg \min_{t \in G} J_{LRE}(\{P'_{i|ic}\}; h'_{ij}) \quad (\text{Eq. 8})$$

3.6 DRIVE database

In this paper we used the images included in the well-known DRIVE database (<http://www.isi.uu.nl/Research/Databases>) to implement the proposed segmentation method and to assess its performance. The DRIVE database contains 40 color retinal images of size 565×584 pixels. The images have been divided into 2 sets, a training set and a test set. Each one contains 20 color retina images. Each set also contains the corresponding segmented images, which were graded by two experts, resulting in sets A and B.

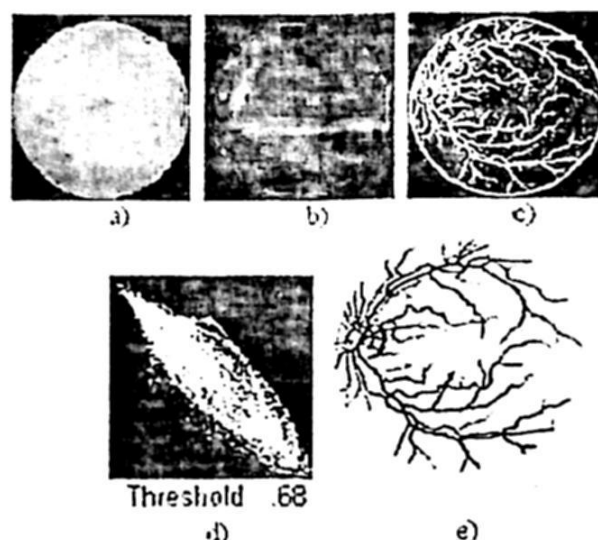


Fig. 3. Steps of the segmentation process for a typical image.

Figure 3 illustrates the results of each step of the pre-processing stage. In figure 3 b) the green color band extracted from the figure 3 a) is shown. The enhanced vessel after applying the BSGMF is presented in figure 3 c). Figure 3 d) illustrates the corresponding co-occurrence matrix and the threshold value obtained for this image depicted as a red cross overlapped in the co-occurrence matrix. Finally, in figure 3 e) the segmented blood vessels using the 2nd local entropy thresholding method is presented. The experimental results show that the proposed segmentation method performs well in extracting vessels, achieving the highest score from all the methods that were compared. There are several parameters of the algorithm that have effects in the performance of the vessel segmentation method. The most significant parameter is the thresholding value. Since the proposed segmentation method obtains automatically this value for each image, it is not necessary to establish a range of thresholding values, and also, it is not necessary the interaction of the user that adjust this value depending on the image case.

3.7 Feature extraction

There are two types of features used for biometric coding, i.e., block features and point wise features [14]. The American National Standards Institute-National Institute

of Standard and Technology (ANSI-NIST) proposed a minutiae-based biometric representation. It includes minutiae location and orientation. Minutia orientation is defined as the direction of the underlying segment at the minutia location. Retinal vessel landmarks are bifurcation, crossings and ending points. Among these features, bifurcations are the most reliable and abundant feature in fundus images [22]. The retinal bifurcations and ending points are unique for each individual and therefore they are useful for a successive process of personal identification [13].

3.7.1.1. Extracting vasculature skeleton image of vessel structure

As can be seen in figure 4 e), the width of the extracted blood vessels is not the same for the entire vasculature. To overcome this, retinal vessel thinning is usually implemented via morphological operation which reduces the width of vessels to a single pixel width line while preserving the extent and connectivity of the original shape. After the skeleton of the retinal image is computed, extracting the minutiae from the one-pixel-wide vessel tree is a trivial task.

3.7.2. Minutiae detection

Minutiae detection in a retinal vessel skeleton is implemented by scanning the thinned vasculature and counting the *crossing number (cn) between veins and arteries* [14]. The *cn* can be defined as follows:

$$cn(P) = \frac{1}{2} \left(\sum_{i=1}^8 |val(P_{i_{mod8}}) - val(P_{i-1})| \right)$$
 (Eq. 9)

Where P_0, P_1, \dots, P_7 are neighbors of p , $val(p) \in (0,1)$. Based on the calculated cross number *cn* it is possible to classify the point according to the following (Table 1):

Table 1. Classification of landmarks according to the *cn* value.

if <i>cn</i> = 1, it is an ending point	Ending points as vascular landmarks is defined by one connection of the pixel with its eight connected neighbors.
if <i>cn</i> = 2, it is an inner point	A segment is a vessel between 2 successive points of bifurcation or between a bifurcation and an ending point. Generally, the continuous blood vessel without vascular landmarks has two connections of its eight connected neighbors [32].
if <i>cn</i> = 3, it is a bifurcation point	Bifurcations consist of a center location that is met by three blood vessel branches [32].
if <i>cn</i> = 4, it is a crossover point	A crossing is an image point which is the intersection of four line segments. Crossings of vessel segments are, for practical purposes, always between a vein and an artery (i.e., crossing between arteries and arteries or between veins and veins are, for practical purposes, non-existent).

3.8. Pruning

Although there are a lot of minutiae detection algorithms available in the literature, minutiae detection accuracy can not reach 100%. In this work it is used a set of simple heuristic rules to eliminate false minutiae.

3.9. Minutiae attribute estimation

As stated before, most common features in biometric applications are minutiae. A retinal image is pre-processed and minutiae of the vascular tree are extracted and coded using their attributes like location and orientation of the vessel they are located on. Then, a list or a graph of minutiae is formed. For each detected minutiae, the following parameters are recorded:

- 1) **Location attributes:** x-coordinate, y-coordinate.
- 2) **Angular attributes:** Orientation, which is defined as the local vessel orientation of the associated vessel.
- 3) **The minutiae type**, i.e., bifurcation point (1) or ending point (2).

The machine representation of a biometric is critical to the success of the matching algorithm. A minimal representation of a processed retinal image is a set $\{(x_i, y_i, ?_i)\}$ of minutiae, i.e., a set of points (x_i, y_i) expressed in some coordinate systems with a vessel direction at this point $?_i$.

4 Experimental results

Figure 4 shows the results of the minutiae attribute estimation stage illustrated for each kind of minutia. In figure 4a) the location of the final detected bifurcation points is indicated using magenta square overlapped to the skeletonized vessel tree, and the corresponding orientation of its three connected segment vessels that conform the bifurcation are indicated using a red line going from the center of the bifurcation point through its three surrounding branches. In figure 4b) the position of the ending points are indicated using blue circles overlapped to the one-pixel-wide vessel tree, and its corresponding orientation are also presented using a red line overlapped to the ending neighbor vessel. Finally, the figure 4c) illustrates the cloud of the entire vascular landmarks and its estimated attributes using the same color code described before for each kind of landmark and for the corresponding orientation attribute.

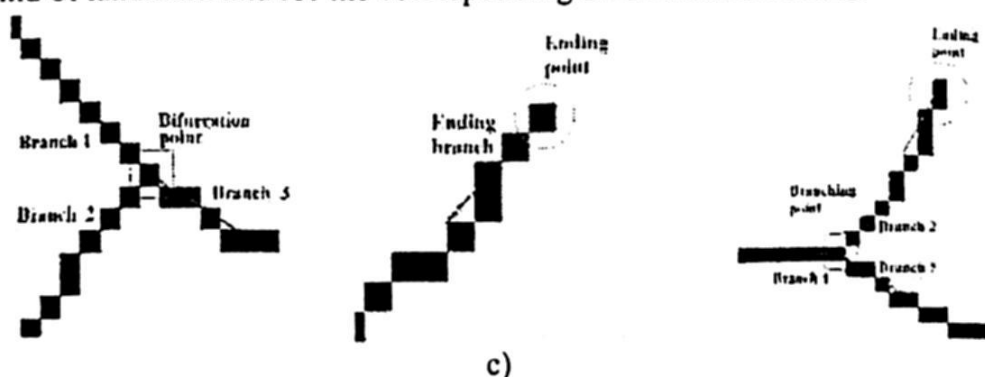


Fig. 4 Landmarks attribute detected in a zoomed zone for a particular minutiae point.

Finally, the characteristic vector can be constructed using all this information, i.e., the landmark position (x, y), the angular landmark attribute (branch orientation) and landmark type (bifurcation point or ending point). All this information integrates the representation used in biometric applications.

4.1. Execution time

How fast the system takes a decision about the claimed identity is of course an important parameter of a biometric system. For this reason, in this work we also analyze the time that the proposed algorithm takes for complete the blood vessel segmentation and the vascular landmarks extraction. On a Pentium (R) Dual-Core T4200 @ 2GHz and 4 GB of internal memory, and with a MATLAB 7.4.0 (R2007a) implementation, it takes in an image of size 565×584 pixels, an average of 4 seconds to obtain the "eye print" from the retinal input image.

5 Conclusions

In this paper it is presented a fast, efficient and automatic minutiae-based algorithm for segmenting retinal vessel tree and extracting its vascular landmarks to use this information as a unique feature for biometrical applications. The proposed method consists of three main steps.

For the pre-processing step we can conclude that:

- a) The extraction of the green color band let us to process more efficiently the information contained in the image and thus reduces the noise effects. The required time is also reduced because only one band is processed.
- b) The mask generation stage uses a thresholding with a free parameter empirically chosen such that pixels with intensity value above that threshold are considered to belong to the Region of Interest (ROI). The threshold is applied to the selected green color band of the image.
- c) Image denoising, correction of uneven illumination and contrast enhancement are needed before applying the vessel segmentation method for landmarks extraction. Uneven illumination (also called shading) is presented in retinal images and must be suppressed in order to achieve more accurate segmentation of the blood vessels. Normalization, correction of the brightness and contrast enhancement are applied in this preprocessing step. Higher contrast between background and vessels in the image, and the small bright noise is removed while most of the capillaries are preserved.

For the main processing step the conclusions are:

- d) The co-occurrence matrix computation gives a powerful tool to obtain an automatic threshold value to segment the vessel tree depending only on the information contained in the image.
- e) Accurate segmentation of retinal images influences directly the performance of minutiae extraction. If more background areas are included in the segmented

retinal image, more false features are introduced; if some parts of the foreground are excluded, useful feature points may be missed. The method is non-supervised, fast and offers high segmentation accuracy.

- f) It is usually desirable to reduce the images to thin representations located along the approximated middle of the original curve or line. Thinning is the process of reducing a shape to its core components while retaining the essential features of the original object.
- g) For the landmarks extraction stage and based on the skeleton image it is possible to extract the retinal vessel landmarks, exploiting the skeleton unitary depth. Using a window of 3 x 3 with 8 neighbor pixels to the central pixel, it is calculated the cross number around the central point.

And, for the post-processing step it is possible to conclude that:

- h) The presence of undesired segment and broken vessels present in a thinned vessel tree may lead to detect many false positive minutiae. Therefore some heuristics rules are used to pre-process the vascular tree.
- i) For each kind of minutia some attributes are estimated. These are: landmark position, orientation and type. Using these features it is possible to create the retinal feature vector.

The average time required to segment and extracts the vascular landmarks information is 4 seconds. Based on these results we consider that our method offers a good alternative for biometrics applications where the time of analysis plays an important role.

References

1. Jain A., Bolle R. and Pankanti S., (2004), *Introduction To Biometric Recognition*, Michigan State University, East Lansing, MI, IBM T. J. Watson Research Center, Yorktown Heights, NY.
2. Simon C., and Goldstein I., (1935), A new scientific method of identification. *New York State. J. Medicine*, 35(18):901-906.
3. Tower, P., (1955), The fundus oculi in monozygotic twins: Report of six pairs of identical twins. *Arch. Ophthalmol.*, 54:225-239.
4. Bruce L. Golden, Bernard E. Rollin, Ralph V. Switzer JR., (2004), Apparatus and method for creating a record using biometric information, U.S. Patent No. 028343.
5. Hill R. B., (1978), Apparatus and method for identifying individuals through their retinal vasculature patterns, U.S. Patent No. 4109237.
6. Hill R. B., (1992), *Retina Identification*, Portland, OR, USA.
7. Marshall J. and Usher D., (2006), Method for generating a unique and consistent signal pattern for identification of an individual, U.S. patent No. 6993161.
8. Derakhshani R. and Ross A., (2007), A Texture-Based Neural Network Classifier for Biometric Identification using Ocular Surface Vasculature, Appeared in *Proc. Of International Joint Conference on Neural Networks (IJCNN)*, Orlando, USA.
9. Usher D., Tosa Y. and Friedman M., (2007), Ocular Biometrics: Simultaneous Capture and Analysis of the Retina and Iris, *Advances in Biometrics Sensors, Algorithms and Systems*, pp. 133-155.

10. http://www.absoluteastronomy.com/topics/Retinal_scan.
11. Usher D. B., (2003), Image analysis for the screening of diabetic retinopathy. PhD thesis, University of London.
12. Jung E. and Hong K., (2006), Automatic Retinal Vasculature Structure Tracing and Vascular Landmark Extraction from Human Eye Image, Proceedings of the International Conference on Hybrid Information Technology, IEEE Computer Society.
13. Bevilacqua V., Cambó S., Cariello L. and Mastronardi G., (2007), Retinal Fundus Hybrid Analysis Based on Soft Computing Algorithms, Communications To Simai Congress, ISSN 1827-9015, Vol. 2.
14. Wu C., (2007), Advanced Feature Extraction Algorithms for Automatic Fingerprint Recognition Systems, a Dissertation submitted to the Faculty of the Graduate School of State University of New York at Buffalo in Particular fulfillment of the requirements for the degree of Doctor of Philosophy.
15. Chaudhuri S., Chatterjee S., Katz N., Nelson N. and Goldbaum M., (1989), Detection of Blood Vessels in Retinal Images Using Two-Dimensional Matched Filters, IEEE Transactions on Medical Imaging, 8(3):263–269.
16. Zhang Y. F. and Zhang Y., (2006), Another Method of Building 2D Entropy to Realize Automatic Segmentation, International Symposium on Instrumentation Science and Technology; Journal of Physics: Conference Series 48, 303–307.
17. Haralick, R.M., Shanmugam, K., and Dinstein, I., (2008), Textural features for image segmentation, IEEE Trans. Syst. Man Cybern., 973, SMC-3, (6), pp. 610–621, 1973.
18. Gonzalez, R., and Woods, J., (2004), Digital image processing, Addison-Wesley.
19. Kullback, S., Information theory and statistics, Communications and Information Theory, Vol. 1, Issue 4, Pp. 417 – 528.
20. Pal N.R. and Pal S.K., (1989), Entropic thresholding, Signal Process 16, 97–10.
21. Zana F. and Klein J. C., (1997), Robust Segmentation of Vessels from Retinal Angiography. In International Conference on Digital Signal Processing, pages 1087–1091, Santorini, Greece.
22. Choe T. E., Cohen I., Lee M. and Medioni G., (2006), Optimal Global Mosaic Generation from Retinal Images, the 18th International Conference on Pattern Recognition.
23. M. Martinez M., Hughes A., Stanton A., Thom S., Bharath A., and Parker K., (1999), Scale-space analysis for the characterization of retinal blood vessels, Medical Image computing and Computer-Assisted Intervention-MICCAI'99, C. Taylor and A. Colchester, eds., pp. 90–97.
24. Staal J., Abramoff M., Niemeijer M., Viergever M. and van Ginneken B., (2004), Ridge-based Vessel segmentation in color images of the retina, IEEE Trans. Med. Imag. 23, 501–509.
25. Nagasubramanian S. and Weale R. A., (2004), Ethnic variability of the vasculature of the optic disc in normal and in glaucomatous eyes. Eur. J. Ophthalmol. 14,(6) pp. 501–507.
- A. Arakala A., Horadam K. J. and Boztas S., (2008), Practical Considerations for Secure Minutiae Based Templates", Proc. 2008 Biometrics Symposium, Tampa, Florida, 23–25, IEEE Press.
26. Soares J.V.B., Leandro J.J.G., Cesar R.M., Jelinek H.F. and Cree M.J., (2006), Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification, IEEE Transactions on Medical Imaging, 25:1214–1222.
27. Wang L. and Bhalerao A., (2003), Model Based Segmentation for Retinal Fundus Images. In Proc. of Scandinavian Conference on Image Analysis (SCIA).
28. Chanwimaluang T. and Fan G., (2003), An efficient blood vessel detection algorithm for retinal images using local entropy thresholding. In: Proc. of the IEEE Intl. Symp. on Circuits and Systems.

29. Jiang X. and Mojon D., (2003), Adaptive local thresholding y verification-based multi-threshold probing with application to vessel detection in retinal images. *IEEE Trans. Pattern Recogn. Anal. Mach. Intell.* 25, 131-137.
30. Siddalingaswamy P.C. and Prabhu G.K., (2007), Automated Detection of Anatomical Structures in Retinal Images, *International Conference on Computational Intelligence and Multimedia Applications* vol. 3, pp.164-168.
31. Rohrschneider K., (2004), Determination of the location of the fovea on the fundus. *Invest. Ophthalmol. Vis. Sci.* 45, 9 pp. 3257-8.
32. Srivastavas V., (2005), Performance of micro-calcification detection algorithms, Master Thesis, Department of Electrical and Computer Engineering.
33. Yang C.W., Dye M., Shuenn C., Chuin W., Chia W., Chien L., Pau C. and Chein C., (2000), Computer-aided diagnostic detection system of venous beading in retinal images, *Society of Photo-Optical Instrumentation Engineers.* 39(5): 1293-1303.

Implementation of a swarm intelligence algorithm to a mobile device

L.E. Gomez¹, J.F. Jimenez¹, J.H. Sossa¹, F.J. Cuevas², O. Pogrebnyak¹, R. Barron¹

¹ Centro de Investigación en Computación-IPN, Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n and M. Othón de Mendizábal, Zacatenco, México, DF. 07738, Mexico

² Centro de Investigaciones en Óptica A.C. Loma del Bosque #115, Col. Lomas del Campestre C.P. 37150, León Gto. México
sgomezb08@sagitario.cic.ipn.mx, jfvielma@cio.mx, hsossa@cic.ipn.mx, fjcuevas@cio.mx, olek@cic.ipn.mx, rbarron@cic.ipn.mx

Abstract. In this paper, we propose the implementation of the algorithm of particle swarm optimization to the mobile platform, with its limitations even in our day is a new way to launch applications with bio-inspired techniques for this type of platform. The particle swarm optimization (PSO) proposed by Kennedy and Eberhart in 1995, is discussed in a numerical optimization of a three benchmark functions taken from the literature. PSO is motivated by social behavior to organisms, such as meetings among birds or fish. Each particle consists of three main parts, its speed, cognitive knowledge and social knowledge.

Keywords: Particle Swarm Optimization; Mobile device; JAVA.

1 Introduction

The optimization in the sense of finding the best solution, or at least a good enough solution for a problem is a field of vital importance in real life. We are constantly solving small problems of optimization, such as the shortest way to get from one place to another, the organization of a book, etc.. In general they are small enough and can be resolved without recourse to external elements to our brain. But as they get larger and more complex, the use of computers to its resolution is unavoidable.

Due the great importance optimization problems over the history computing, have developed multiple methods try to solve them. Around the seventies came a class of algorithms are not accurate, whose basic idea was combine different heuristics to a higher level to get an exploration of the search space efficiently and effectively. These techniques have been called metaheuristics.

From the different descriptions of metaheuristics which are found in the literature can be rankings were certain fundamental properties that characterize this type method:

- The metaheuristic are templates by general strategies or "guide" the search process.

- The goal is an exploration of the search space efficiently to find solutions (almost) optimal.
- The metaheuristic algorithms are not exact and are generally nondeterministic.
- They may incorporate mechanisms to avoid areas of non-optimal search space.
- The basic layout any metaheuristic is general and not dependent on the problem to be solved.
- The functions used in metaheuristics is goodness (fitness functions) to quantify the appropriateness of a particular solution

Summarizing these points, a metaheuristic is a high-level strategy that uses different methods to explore the search space.

With the dramatic increase and sophistication of mobile devices such as cell phones, also comes the demand for applications that run on them. corporations want to expand consumer devices for mobile communications devices for voice communications applications traditionally found on laptops and PC's.

Developers, handset manufacturers, are eager to fill this need, however there is a serious difficulty, mobile communications devices use different application platforms and operating systems, in addition to the various virtual machines which do not have all the libraries in J2SE (Java 2 Standard Edition).

An application that runs on a device hardly runs in another. Mobile devices lack a standard application platform and the same operating system, which causes the development of applications for mobile devices, is a financial risk for developers.

The lack of standards is nothing new to the area of computer technology or any new development. Traditionally, hardware device manufacturers try to force the market to accept its standards. Owners of as standards in the industry, other times, industry leaders formed a consortium, such as Java Community Process Program, to collectively define a standard [1].

The J2ME architecture is oriented to small devices and embedded systems such as mobile phones, PDAs, Pockets, etc. In order have a J2ME runtime environment that meets the requirements of a wide range of devices and target markets is required to be made to Figure 1:

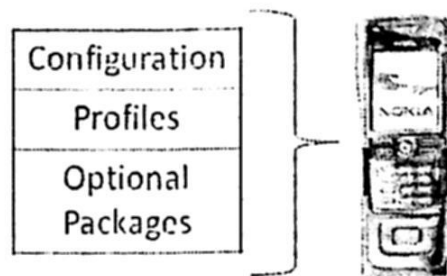


Fig. 1. Mobile Device Architecture.

The settings consist of a virtual machine and a minimal set of function libraries. Provide basic functionality for a set of devices that share similar characteristics, such as memory management or network connectivity. There are two types of configurations such as CLDC devices aimed at processing and memory constraints, and the CDC focused on devices with more resources.

To form a complete runtime environment targeted to a category of devices, the settings have to be combined with a set of APIs to a higher level, called profiles, which go a step further in defining the life cycle model of applications, user interface and access to specific properties of the devices. At present there are four profiles: MIDP (Mobile Information Device Profile), FP (Foundation Profile), PP (Personal Profile) and PBP (Personal Basis Profile) [2].

Regarding J2ME optional packages can be extended by combining various optional packages with CLDC and CDC along with their profiles. These packages were created to meet specific requirements and offer a set of standard APIs for using both existing and emerging technologies such as Bluetooth, Web services, wireless messaging, multimedia capabilities and connectivity to databases. Because they are modular, manufacturers can incorporate according needed to improve the features supported.

There are two versions of MIDP: 1.0 and 2.0. In MIDP, often uses the term to refer a Midlet application that supports the profile. For distribution, one or more MIDlets are grouped in bundles called suites, which consist of a file format Jar and an optional descriptor Jad [3], this is shown in Figure 2.

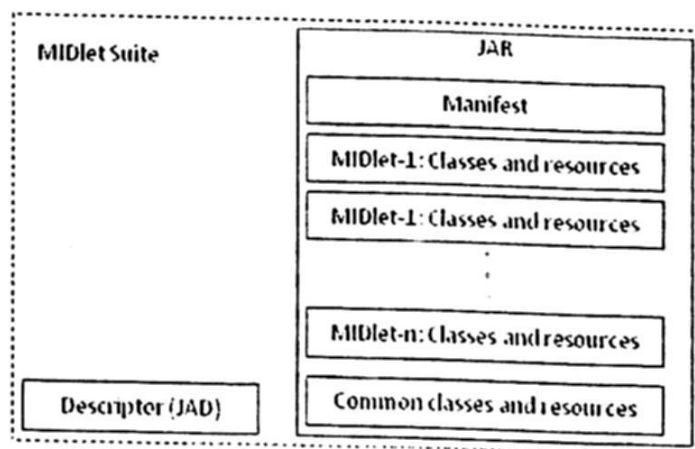


Fig. 2. Structure of a MIDlet suite.

The active state of a MIDlet is when the application is running. State is destroyed when the application terminates and frees the memory required. The paused state occurs when the application stops for a particular event or has not been activated yet, when it created [4], this is shown in Figure 3.

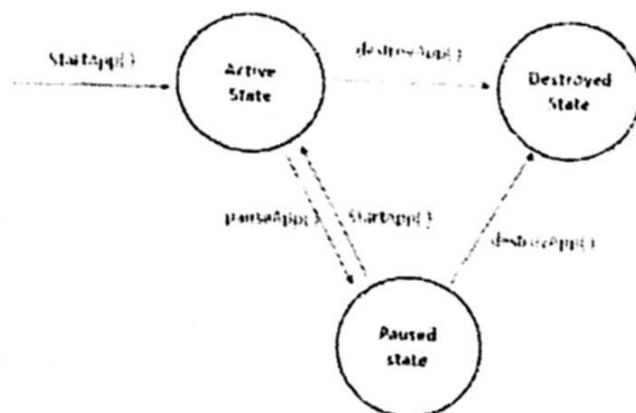


Fig. 3. Life cycle of a MIDlet.

2 Particle Swarm Optimization (PSO)

PSO is a stochastic global optimization method which is based on simulation of social behavior. As in GA and ES, PSO exploits a population of potential solutions to probe the search space. In contrast to the aforementioned methods in PSO no operators inspired by natural evolution are applied to extract a new generation of candidate solutions. Instead of mutation PSO relies on the exchange of information between individuals, called *particles*, of the population, called *swarm*. In effect, each particle adjusts its trajectory towards its own previous best position, and towards the best previous position attained by any member of its neighborhood [5]. In the global variant of PSO, the whole swarm is considered as the neighborhood. Thus, global sharing of information takes place and particles profit from the discoveries and previous experience of all other companions during the search for promising regions of the landscape. To visualize the operation of the method consider the case of the single objective minimization case: promising regions in this case possess lower function values compared to others, visited previously.

Several variants of PSO have been proposed up to date, following Eberhart and Kennedy who were the first to introduce this method [6], [7], [8]. The variants which were applied in our experiments are exposed in the following paragraphs.

Table 1 shows a number of terms used in the traditional PSO algorithm.

Table 1. Terms of PSO algorithm.

Term	Description
Particle / Agent	An individual of the swarm
Location / Position	Coordinates of an agent in N-dimensional space represents a solution to the problem
Swarm	A whole collection of agents / A population of individuals
<i>Fitness</i>	A number that indicates the quality of a given solution (represented by a location in the solution space)
<i>pbest</i> (Personal best)	It is the best location obtained by a given agent during the process
<i>gbest</i> (Global best)	It is the best location in all the particle swarm

Initially, let us define the notation adopted in this paper: assuming that the search space is D dimensional, the i -th particle of the swarm is represented by a D dimensional vector $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ and the best particle of the swarm, i.e. the particle with the lowest function value, is denoted by index g . The best previous position (i.e. the position corresponding to the best function value) of the i -th particle is recorded and represented as $P_i = (p_{i1}, p_{i2}, \dots, p_{iD})$, and the position change (velocity) of the i -th particle is $V_i = (v_{i1}, v_{i2}, \dots, v_{iD})$.

The particles are manipulated according to the following equations (the superscripts denote the iteration):

$$V_i^{k+1} = \gamma(\omega V_i^k + c_1 \text{rand}_1^k() (P_i^k - X_i^k) + c_2 \text{rand}_2^k() (P_g^k - X_i^k)) \quad (1)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1}, \tag{2}$$

where $i = 1, 2, \dots, N$, and N is the size of the population; γ is a *constriction factor* which is used to control and constrict velocities; ω is the *inertia weight*; c_1 and c_2 are two positive constants, called the cognitive and social parameter respectively; $rand_{i1}()$ and $rand_{i2}()$ are random numbers uniformly distributed within the range $[0, 1]$. Eq. (1) is used to determine the i -th particle's new velocity, at each iteration, while Eq. (2) provides the new position of the i -th particle, adding its new velocity, to its current position. The performance of each particle is measured according to a fitness function, which is problem dependent. In optimization problems, the fitness function is usually identical with the objective function under consideration.

The update of a particle in general is illustrated in Figure 4.

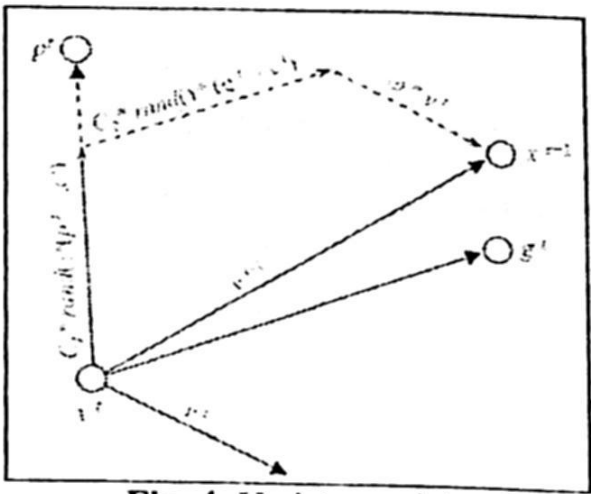


Fig. 4. Update particle.

The role of the inertia weight ω is considered important for the PSO's convergence behavior. The inertia weight is employed to control the impact of the previous history of velocities on the current velocity. Thus, the parameter ω regulates the tradeoff between the global (wide ranging) and the local (nearby) exploration abilities of the swarm. A large inertia weight facilitates exploration (searching new areas), while a small one tends to facilitate exploitation, i.e. fine tuning the current search area. A proper value for the inertia weight ω provides balance between the global and local exploration ability of the swarm, and, thus results in better solutions. Experimental results imply that it is preferable to initially set the inertia to a large value, to promote global exploration of the search space, and gradually decrease it to obtain refined solutions [9]. The initial population can be generated either randomly or by using a Sobol sequence generator [10], which ensures that the D -dimensional vectors will be uniformly distributed within the search space.

The PSO technique has proven to be very efficient for solving real valued global unconstrained optimization problems [11],[12]. In the next section experimental results of the performance of PSO in mobile device.

3 Experiment

Three numerical optimization problems were chosen to compare the relative performance of PSO algorithm in a mobile device. These functions are standard functions of all test patterns and minimization problems.

3.1 Functions

The functions are unimodal. All functions are designed to have global minimum near the origin.

The first test function is the function given by equation Sphere:

$$\begin{aligned} f_1(x) &= \sum_{i=1}^n x_i^2 \\ -100 &\leq x_i \leq 100 \\ \min(f_1) &= f_1(0, \dots, 0) = 0 \end{aligned} \quad (3)$$

x is a real vector of dimension n and x_i is the i -th element in the vector. The results of optimizing the function (3), PSO heuristics are shown in Table 2.

Table 2 Results from the function (3).

Iterations	Particles	Velocity	Variables	Fitness
500	50	4	10	6.15
400	100	4	10	2.88
300	50	4	10	2.03
300	30	4	10	3.19
200	100	4	10	5.43

The second function is Schwefel's problem, given by the equation:

$$\begin{aligned} f_2(x) &= \sum_{i=1}^n |x_i| + \prod_{i=1}^n |x_i| \\ -10 &\leq x_i \leq 10 \\ \min(f_2) &= f_2(0, \dots, 0) = 0 \end{aligned} \quad (4)$$

The results of optimizing the function (4), PSO heuristics are shown in Table 3.

Table 3 Results from the function (4).

Iterations	Particles	Velocity	Variables	Fitness
400	30	1	10	1.26
300	50	1	10	1.15
300	30	1	10	1.27
250	50	1	10	1.25
200	30	1	10	1.49

The third function is Step, given by equation:

$$f_3(x) = \sum_{i=1}^n (x_i + 0.5)^2$$
$$-100 \leq x_i \leq 100$$
$$\min(f_3) = f_1(0,...,0) = 0$$

(5)

The results of optimizing the function (5), PSO heuristics are shown in Table 4.

Table 4 Results from the function (5).

Iterations	Particles	Velocity	Variables	Fitness
600	20	4	10	0.0
500	20	4	10	0.0
450	20	4	10	0.0
400	20	4	10	3.0
350	20	4	10	3.0

3.2 PSO Simulator

The PSO was programmed in the Java language in its version J2ME for programming with IDE I use Netbeans 6.8 and Sun Java Wireless Toolkit 2.5.2 for CLDC, the first tests were in a simulator on a PC, these performances are shown in Figure 5.

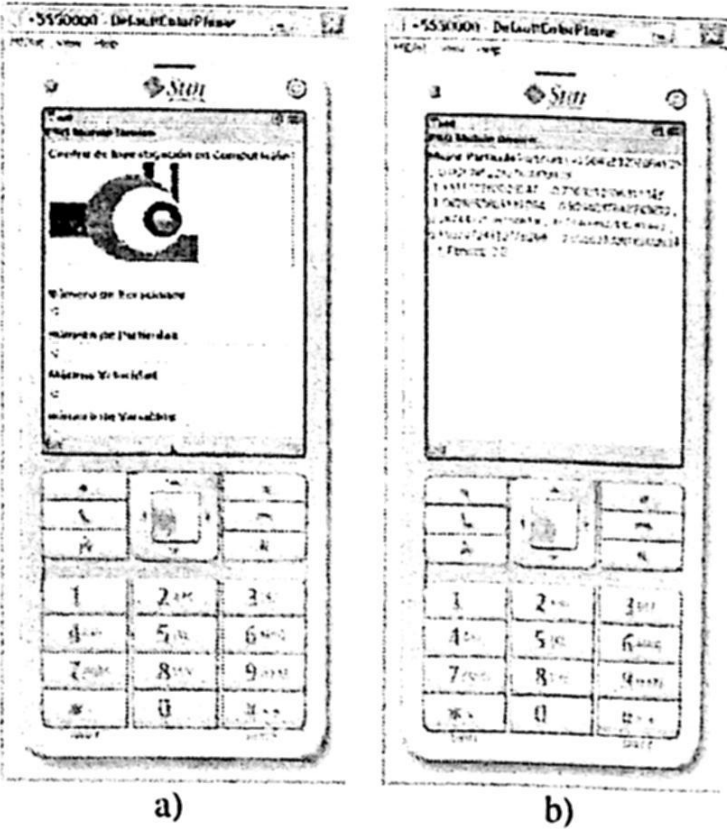


Fig. 5. a) Home Screen Simulator, b) Results Screen Simulator.

2.1 PSO Mobile Devices

To run on mobile devices requires that the device has JAVA support, the tests presented here in two phones, a Nokia brand, model N91, the other brand W760i SONY ERIKSSON, executions shown in Figures 6 and 7.

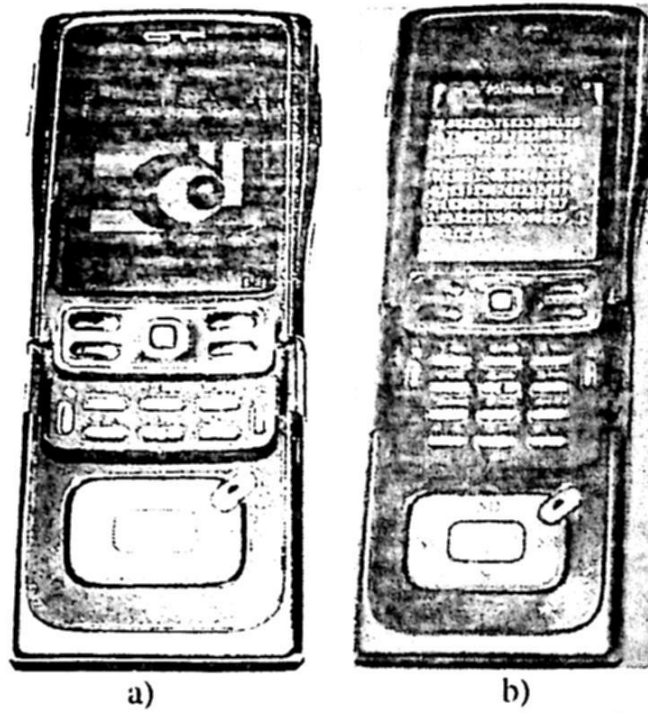


Fig. 6. a) Start Screen in Nokia N91, b) Results Screen in Nokia N91.

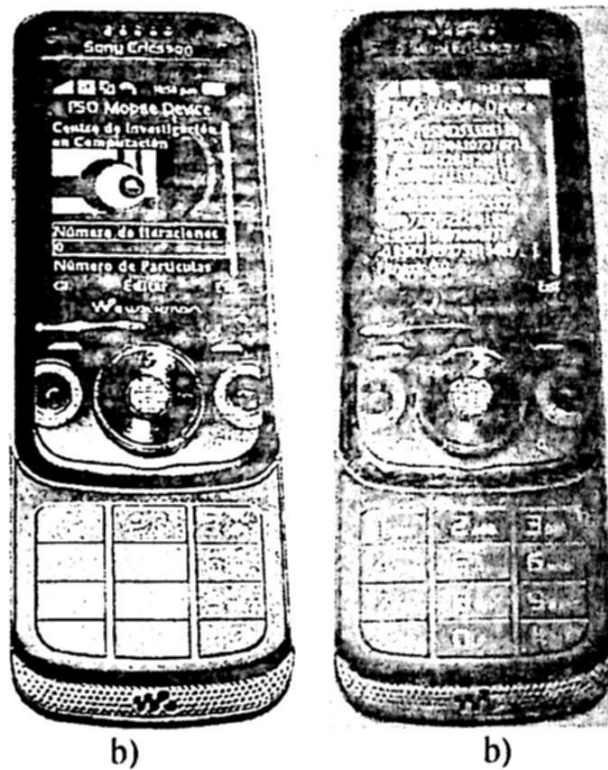


Fig. 7. a) Start Screen in Sony Ericsson w760i, b) Results Screen in Sony Ericsson w760i.

4 Conclusions

In this paper, it was possible the analysis of three well-known problems in the literature. The results showed that with PSO optimized mobile devices programmatically gives very good results to minimize such functions, which have been tested with other techniques such as bio-inspired Differential Evolution.

The implementation of bio-inspired techniques to a new platform such as mobile devices opens a further step in the investigation. For many years the technology has advanced so amazing, just look around us to realize, before he took on a Palm, pocket, and mobile phone was to please a taste, now cover human needs. Support multiple applications, ranging from taking a picture, listening to music, to do complex numerical computations, important for the realization of this project.

Programming for mobile now has endless barriers, since they do not have all the benefits of programming language when it is for PC's, talking specifically about the mobile libraries included. However, there are ways to get the results from its own library of programming for that platform. The program was tested on different mobile devices.

Acknowledgements. We wish to thank the Centro de Investigación en Computación of the I.P.N. by the support to accomplish this project. L.E. Gomez and J.F. Jimenez thanks CONACYT by the scholarship received to complete his doctoral studies. H. Sossa thanks the SIP-IPN under grant 20091421 for the support. R. Barron thanks the SIP-IPN under grant 20100379 for the support. O. Probegnyak thanks the SIP-IPN for the support. H. Sossa also thanks CINVESTAV-GDL for the support to do a sabbatical stay from December 1, 2009 to May 31, 2010. Authors thank the European Union, the European Commission and CONACYT for the economical support. This paper has been prepared by economical support of the European Commission under grant FONCICYT 93829. The content of this paper is an exclusive responsibility of the CIC-IPN and it cannot be considered that it reflects the position of the European Union. Finally, authors thank the reviewers for their comments for the improvement of this paper.

References

- [1] Microjava.: Introducción a J2ME y KVM, Tutorial, <http://microjava.com>.
- [2] Keogh James.: The complete reference, Edit. McGraw-Hill, 2003.
- [3] Qusay Mahmoud.: Learning wireless Java, Edit. O'Really, December 2001.
- [4] Gálvez Rojas Sergio., Ortega Díaz Lucas.: Java a tope: J2ME. Depto de lenguajes y ciencias de la computación.
- [5] Kennedy, J.: The Behavior of Particles. *Evol. Progr.* VII (1998) 581-587.
- [6] Eberhart, R.C., Simpson, P.K., Dobbins, R.W.: *Computational Intelligence PC Tools*. Academic Press Professional, Boston (1996).
- [7] Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. *Proc. IEEE Int. Conf. Neural Networks*. Piscataway, NJ (1995) 1942-1948.
- [8] Kennedy, J., Eberhart, R.C.: *Swarm Intelligence*. Morgan Kaufmann (2001).

- [9] Shi, Y., Eberhart, R.C.: Parameter Selection in Particle Swarm Optimization. *Evolutionary Programming VII* (1998) 591-600.
- [10] Press, W.H., Vetterling, W.T., Teukolsky, S.A., Flannery, B.P.: *Numerical Recipes in Fortran 77*. Cambridge University Press, Cambridge (1992).
- [11] Parsopoulos, K.E., Plagianakos, V.P., Magoulas, G.D., Vrahatis, M.N.: Objective Function "Stretching" to Alleviate Convergence to Local Minima. *Nonlinear Analysis TMA* 47(5) (2001) 3419-3424.
- [12] Parsopoulos, K.E., Vrahatis, M.N.: Initializing the Particle Swarm Optimizer Using the Nonlinear Simplex Method. A. Grmela, N.E. Mastorakis (eds.), *Advances in Intelligent Systems, Fuzzy Systems, Evolutionary Computation*. WSEAS Press(2002) 216-221.

Demodulation of a single Interferogram by use a Parametric Method based on a Differential Evolution

J.F. Jimenez¹, F.J. Cuevas², J.H. Sossa¹, L.E. Gomez¹

¹ Centro de Investigación en Computación-IPN, Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n and M. Othón de Mendizábal, Zacatenco, México, DF. 07738, Mexico

² Centro de Investigaciones en Óptica A.C. Loma del Bosque #115, Col. Lomas del Campestre C.P. 37150, León Gto. México

jfvielma@cio.mx, fjcuevas@cio.mx, hsossa@cic.ipn.mx, sgomez08@sagitario.cic.ipn.mx

Abstract. A parametric method to carry out fringe pattern demodulation by means of Differential Evolution is presented. The phase is approximated by the parametric estimation of an n th-grade polynomial so that no further unwrapping is required. On the other hand, a different parametric function can be chosen according to the prior knowledge of the phase behavior. A differential evolution is codified with the parameters of the function that estimates the phase. A fitness function is established to evaluate the vectors, which considers: (a) the closeness between the observed fringes and the recovered fringes, (b) the phase smoothness, (c) the prior knowledge of the object as its shape and size. The differential evolution evolves until a fitness average threshold is obtained. The method can demodulate noisy fringe patterns and even a one-image closed-fringe pattern successfully.

Keywords: Phase retrieval; Fringe analysis; Optical metrology; Differential Evolution.

1 Introduction

In optical metrology, a fringe pattern (interferogram) can be represented using the following mathematical expression:

$$I(x, y) = a(x, y) + b(x, y) \times \cos(\omega_x x + \omega_y y + \phi(x, y) + n(x, y)) \quad (1)$$

where x, y are integer values representing indexes of the pixel location in the fringe image, $a(x, y)$ is the background illumination, $b(x, y)$ is the amplitude modulation and is $\phi(x, y)$ the phase term related to the physical quantity being measured. ω_x and ω_y are the angular carrier frequency in directions x and y . The term $n(x, y)$ is an additive phase noise. The purpose of any interferometric technique is to determine the phase term, which is related to the physical quantity, being measured. One way to calculate the phase term $\phi(x, y)$ is by using the phase-shifting technique (PST) [1–5], which needs at least three phase-shifted interferograms. The phase shift among

interferograms must be known and experimentally controlled. This technique can be used when mechanical conditions are met throughout the interferometric experiment.

On the other hand, when the stability conditions mentioned are not covered, there are many techniques to estimate the phase term from a single fringe pattern, such as: the Fourier method [6,7], the Synchronous method [8] and the phase locked loop method (PLL) [9], among others. However, these techniques work well only if the analyzed interferogram has a carrier frequency, a narrow bandwidth and the signal has low noise. Moreover, these methods fail for phase calculation of a closed-fringe pattern. Additionally, the Fourier and Synchronous methods estimate the phase wrapped because of the arctangent function used in the phase calculation, so an additional unwrapping process is required. The unwrapping process is difficult when the fringe pattern includes high amplitude noise, which causes differences greater than 2π radians between adjacent pixels [10–12].

Recently, regularization [13–15] and neural networks techniques [16,17] have been used to work with fringe patterns, which contain a narrow bandwidth and noise.

In this work, we propose a technique to determine the phase $\phi(x, y)$, from a fringe pattern with a narrow bandwidth and/or noise, by parametric estimation of a global non-linear function instead of local planes in each site (x, y) as it was proposed in [13,18]. Differential Evolution (DE) algorithm is a new heuristic approach mainly having three advantages; Finding the true global minimum regardless of the initial parameter values, fast convergence, and using few control parameters. DE algorithm is a population based algorithm like genetic algorithms using similar operators; crossover, mutation and selection. When a noisy closed fringe pattern is demodulated, neither a low-pass filter nor a thresholding operator is required. On the other hand, regularization techniques need both of them.

2 DE applied to phase recovery

The standard Differential Evolution (DE) algorithm, belonging to the family of Evolutionary Algorithms, was described by Storn and Price [19],[20]. It is based on evolution of a population of vectors, which encode potential solutions to the problem and traverse the fitness landscape by means of genetic operators that are supposed to bias their evolution towards better solutions. DE is a relatively new optimisation technique compared with other more established Evolutionary Algorithms, such as Genetic Algorithms, Evolutionary Strategy, and Genetic Programming [21].

DE is an optimization algorithm that creates new candidate solutions by combining the parent vector and several other vectors of the same population. A candidate replaces the parent only if it has better fitness [21],[22]. DE uses genetic operators, referred to as mutation, crossover and selection. The role of the genetic operators is to ensure that there is sufficient pressure to obtain even better solutions from good ones (exploitation) and to cover sufficiently the solution space to maximize the probability of discovering the global optimum (exploration).

During the initialization of the algorithm, a population of NP vectors, where NP is the number of vectors, each of dimension D (Which is the number of decision

variables in the optimization problem), is randomly generated over the feasible search space.

The fringe demodulation problem is difficult to solve when the level of noise affecting the fringe pattern is elevated, since many solutions are possible even for a single noiseless fringe pattern. Besides, the complexity of the problem is increased when a carrier frequency does not exist (closed fringes are presented).

Given that for a closed fringe interferogram there are multiple phase functions for the same pattern, the problem is stated as an ill-posed problem in the Hadamard sense, since a unique solution cannot be obtained [22]. It is clear that image of a fringe pattern $I(x, y)$ will not change if $\phi(x, y)$ in Eq. (1) is replaced with another phase function $\tilde{\phi}(x, y)$ given by

$$\tilde{\phi}(x, y) = \begin{cases} -\phi(x, y) + 2\pi & (x, y) \in R, \\ \phi(x, y) & (x, y) \notin R \end{cases} \quad (2)$$

where R is an arbitrary region and k is an integer. In this work, a DE is presented to carry out the optimization process, where a parametric estimation of a non-linear function is proposed to fit the phase of a fringe pattern. Then, DE technique fits a global non-linear function instead of a local plane to each pixel just like it is made in regularization techniques [13,18]. The fitting function is chosen depending on the prior knowledge of the demodulation problem as object shape, carrier frequency, pupil size, etc. When no prior information about the shape of $\phi(x, y)$ is known, a polynomial fitting is recommended. In this paper, authors have used a polynomial fitting to show how the method works.

The purpose in any application of DE is to evolve a population of size NP (which codifies NP possible solutions to the problem) using mutation, crossover and selection of each vector, with the goal of optimizing a fitness function adequate to the problem to solve.

In this work, the fitness function U , which is used to evaluate the p th vector a^p in the population, is given by

$$U(a^p) = \alpha - \sum_{y=1}^{R-1} \sum_{x=1}^{C-1} \left\{ \left(I_N(x, y) - \cos(\omega_x x + \omega_y y + f(a^p, x, y)) \right)^2 + \lambda \left[\left(f(a^p, x, y) - f(a^p, x-1, y) \right)^2 + \left(f(a^p, x, y) - f(a^p, x, y-1) \right)^2 \right] \right\} m(x, y), \quad (3)$$

where x, y are integer values representing indexes of the pixel location in the fringe image. Superindex p is an integer index value between 1 and NP , which indicates the number of vectors in the population. $I_N(x, y)$ is the normalized version of the detected irradiance at point (x, y) . The data were normalized in the range $[-1, 1]$. ω_x and ω_y are the angular carrier frequencies in directions x and y . The Function $f(\cdot)$ is the selected fitting function to carry out the phase approximation. $R \times C$ is the image resolution where fringe intensity values are known and λ is a smoothness

weight factor (it should be clear for the reader that a higher value of parameter λ implies a smoother function to be fitted). The binary mask $m(x, y)$ is a field which defines the valid area in the fringe pattern. The parameter a can be set to the maximum value of the second term (in negative sum term) at Eq. (3) in the first vector population, which is given by

$$\alpha = \max_p \left\{ \sum_{y=1}^{R-1} \sum_{x=1}^{C-1} \left\{ (I_N(x, y) - \cos(\omega_x x + \omega_y y + f(a^p, x, y)))^2 \right. \right. \\ \left. \left. + \lambda \left[(f(a^p, x, y) - f(a^p, x-1, y))^2 \right. \right. \right. \\ \left. \left. \left. + (f(a^p, x, y) - f(a^p, x, y-1))^2 \right] \right\} m(x, y), \right. \quad (4)$$

parameter α is used to convert the proposal from minimal to maximal optimization since a fitness function in a DE is considered to be a nonnegative figure of merit and profit [19].

The first term (in negative sum term) at Eq. (3) attempts to keep the local fringe model close to the observed irradiances in least-squares sense. The second term (in negative sum term) at Eq. (3) is a local discrete difference, which enforces the assumption of smoothness and continuity of the detected phase.

At the beginning of a DE, a set of random solutions are codified in a vector population of size NP . Each vector a is formed by the parameter function vector (possible solution) and chained string such as:

$$a = [a_0 | a_1 | a_2 | \dots | a_H] \quad (5)$$

Each dimension a_i is a random real number in a defined search range $(\min(a_i), \max(a_i))$ (the user defined maximum and minimum of a_i). These values can be initialized using prior knowledge (e.g. in the polynomial case, components x and y are related to the interferogram tilt so if a closed fringe is presented, then these values are near 0). Every dimension is generated as:

$$a_i = \text{random}(\min(a_i), \max(a_i)) \quad (6)$$

Therefore, the population of DE consists of NP D-dimensional parameter vectors $X_{i,G}$, where $i = 1, 2, \dots, NP$, for each generation G .

2.1 Mutation

In the mutation step, a difference between two randomly selected vectors from the population is calculated. This difference is multiplied by a fixed weighting factor, F , and it is added to a third randomly selected vector from the population, generating the mutant vector, $V[1-3]$.

For each target vector $x_{i,G}$, a mutant vector is produced by;

$$v_{i,G+1} = x_{i,G} + K \cdot (x_{r1,G} - x_{i,G}) + F \cdot (x_{r2,G} - x_{r3,G}) \quad (7)$$

where $i, r_1, r_2, r_3 \in \{1, 2, \dots, NP\}$ are randomly chosen and must be different from each other. In Equation (7), F is the scaling factor which has an effect on the difference vector $(x_{r_2,G} - x_{r_3,G})$, K is the combination factor.

2.2 Crossover

After mutation, the crossover is performed between the vector (X) and the mutant vector (V) (Figure 1), using the scheme in (8) to yield the trial vector (U). The crossover probability is determined by the crossover constant (CR), and its purpose is to bring in diversity into the original population [23].

The parent vector is mixed with the mutated vector to produce a trial vector $u_{j,G+1}$

$$u_{j,G+1} = \begin{cases} v_{j,G+1} & \text{if } (rnd_j \leq CR) \text{ or } j = rn_i, \\ q_{j,G} & \text{if } (rnd_j > CR) \text{ or } j \neq rn_i, \end{cases} \quad (8)$$

where $j = 1, 2, \dots, D$; $rnd_j \in [0, 1]$ is the random number; CR is crossover constant $\in [0, 1]$ and $rn_i \in (1, 2, \dots, D)$ is the randomly chosen index, which ensures that $u_{i,G+1}$ gets at least one parameter from $v_{i,G+1}$ [19].

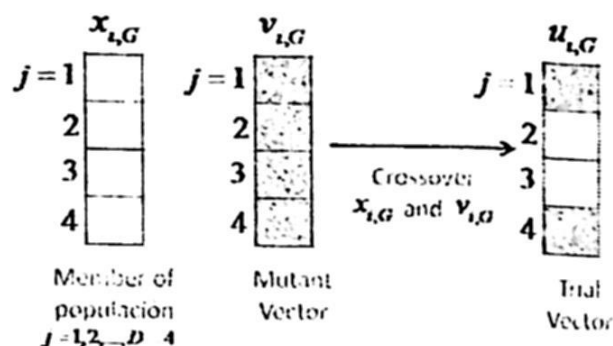


Fig. 1. Illustration of the crossover process for $D=4$.

There are different variants that can be used in mutation and crossover, and they are referred to as DE/x/y/z, where x specifies the vector to be mutated which currently can be "rand" (a randomly chosen population vector) or "best" (vector of the lowest cost from the current population); y is the number of difference vectors used and z denotes the crossover scheme [20].

2.3 Selection

In the last step, called selection, the new vectors (U) replace their predecessors if they are closer to the target vector.

All solutions in the population have the same chance of being selected as parents without dependence of their fitness value. The child produced after the mutation and crossover operations is evaluated. Then, the performance of the child vector and its parent is compared and the better one is selected. If the parent is still better, it is retained in the population.

Figure 2 shows DE's process in detail: the difference between two population members (1,2) is added to a third population member (3). The result (4) is subject to the crossover with the candidate for replacement (5) to obtain a proposal (6). The proposal is evaluated and replaces the candidate if it is found to be better.

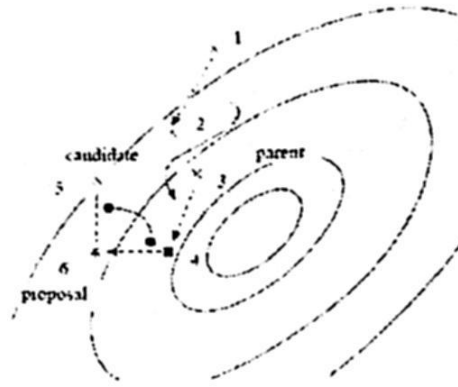


Fig. 2. Obtaining a new proposal in DE.

DE has shown to be effective on a large range of classical optimization problems, and it showed to be more efficient than techniques such as Simulated Annealing and Genetic Algorithms [22],[23]. However, its capability of finding the global optimum is very sensitive to the choice of the control variable F and CR [24]. Consistently with related studies [22],[23],[24], the paper highlights an undesirable behaviour of the algorithm, i.e., the DE does not find the global optimum (value to reach - VTR) when 100% of the population is trapped in a basin of attraction of a local optimum.

2.4 DE convergence

The DE convergence mainly depends on the population size. It should be clear that if we increase the population size, more vectors will search the global optimum and a best solution will be found in a minor number of iterations, although the processing time can be increased [24].

To stop the DE process, different convergence measures can be employed. In this paper, we have used a relative comparison between the fitness function value of the best vectors in the population and value α , which is the maximum possible value to get in Eq. (3). Then, we can establish a relative evaluation of uncertainty to stop the DE as:

$$\left| \frac{\alpha - U(a^*)}{\alpha} \right| \leq \varepsilon, \quad (8)$$

where $U(a^*)$ is the fitness function value of the best vectors in the population in the current iteration, and ε is the relative error tolerance. Additionally, we can stop the process in a specified number of iterations, if Eq. (9) is not satisfied.

3 Experiment

The parametric method using a DE was applied to calculate phase from shadow moiré closed fringe pattern. We used a population size equal to 100, F is calculated by values of "F_lower" and "F_higher", in the ranges $[0.1, 0.9]$. In each vector, the coded coefficients of a fourth degree polynomial were included. The following polynomial was coded in each vector:

$$p_4(x, y) = a_0 + a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2 + a_6x^3 + a_7x^2y + a_8xy^2 + a_9y^3 + a_{10}x^4 + a_{11}x^3y + a_{12}x^2y^2 + a_{13}xy^3 + a_{14}y^4 \quad (9)$$

so that 15 coefficients were configured in each vector inside population to be evolved.

3.1 Close fringe pattern

A low contrasted noisy closed fringe pattern was generated in the computer using the following expression:

$$I(x, y) = 127 + 63 \cos(p_4(x, y) + \eta(x, y)), \quad (10)$$

where

$$p_4(x, y) = -0.7316x - 0.2801y + 0.0065x^2 + 0.00036xy - 0.0372y^2 + 0.00212x^3 + 0.000272x^2y + 0.001xy^2 - 0.002y^3 + 0.000012x^4 + 0.00015x^3y + 0.00023x^2y^2 + 0.00011xy^3 + 0.000086y^4 \quad (11)$$

and $\eta(x, y)$ is the uniform additive noise in the range $[-2\text{radians}, 2\text{radians}]$. Additionally, the fringe pattern was generated with a low resolution of 60×60 . In this case, we use a parameter search range of $[-1, 1]$. The population of vectors was evolved until the number of iterations and relative error tolerance ε was 0.05 in Eq. (9). This condition was achieved in 77s on a AMD Turion X2-2.4 GHz computer. The fringe pattern and the contour phase field of the computer generated interferogram are shown in Fig. 3.

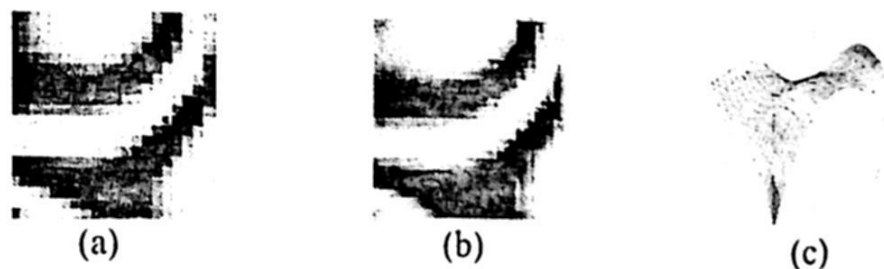


Fig. 3. (a) Original fringe pattern, (b) phase field obtained by using DE technique and (c) phase obtained in 3D.

The DE technique was used to recover the phase from the fringe pattern. The fringe pattern and the phase estimated by DE is shown in Fig. 3. The normalized RMS error was 0.12 radians and the peak-to-valley error was 0.94 radians. Tests are shown on Table 1, the best vectors for the testers are shown on Table 2, and worst vectors for the testers is shown on Table 3.

Table 1. Table of parameters of "F_lower" and "F_higher"

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	2.30E-02	4.84E-01	7.79E-01	6.29E-02	1.64E-01	2.24E-02	2.63E-01	3.77E-01	7.58E-01
0.2	3.60E-02	4.80E-02	7.19E-01	9.19E-01	1.21E+00	1.52E-02	1.28E-03	1.40E-04	9.63E-03
0.3	1.99E-02	4.37E-02	1.08E-03	1.32E+00	1.83E-03	1.12E-04	1.35E+00	8.28E-03	2.24E+00
0.4	7.00E-01	1.04E+00	1.51E+00	3.62E-01	1.58E-04	9.44E-04	1.61E+00	1.92E-03	2.90E+00
0.5	1.03E-03	2.81E-01	1.47E+00	1.67E+00	3.54E-04	1.76E+00	2.12E+00	1.92E+00	1.94E+00
0.6	7.88E-01	2.02E-01	1.44E+00	1.35E+00	1.46E+00	2.23E+00	1.80E+00	2.72E+00	3.14E+00
0.7	3.15E-01	1.19E+00	1.95E+00	1.17E+00	1.88E+00	2.31E+00	2.11E+00	3.29E+00	2.87E+00
0.8	9.20E-01	1.74E+00	1.31E+00	1.91E+00	2.27E+00	2.02E+00	2.11E+00	2.77E+00	3.79E+00
0.9	1.03E-03	1.89E+00	1.40E+00	3.09E-03	2.71E+00	3.11E+00	2.48E+00	2.08E+00	3.07E+00

Table 2. Shows of the best vectors



















F_lower	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
F_higher	0.5	0.3	0.3	0.9	0.4	0.3	0.2	0.2	0.2
Error	1.03E-03	4.37E-02	1.08E-03	3.09E-03	1.58E-04	1.12E-04	1.28E-03	1.40E-04	9.63E-03
									

Table 3. Shows of the worst vectors

F_lower	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
F_higher	0.8	0.9	0.7	0.8	0.9	0.9	0.9	0.7	0.8
Error	9.20E-01	1.89E+00	1.95E+00	1.91E+00	2.71E+00	3.11E+00	2.48E+00	3.29E+00	3.79E+00
									

The phases: original, best vector, as worst vector, is shows in the Fig. 4.

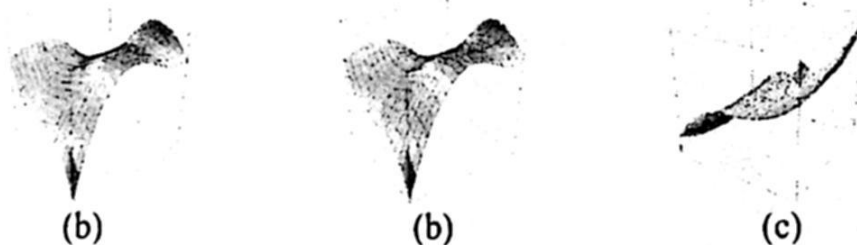


Fig. 4. Phases: (a) Original, (b) best vector of DE technique and (c) worst vector of DE technique.

4 Conclusions

A DE was applied to recover the modulating phase from closed and noisy fringe patterns. A fitness function, which considers the prior knowledge of the object being tested, is established to approximate the phase data. In this work a fourth degree polynomial was used to fit the phase.

A population of vectors was generated to carry out the optimization process. Each vector was formed by a codified string of polynomial coefficients. Then, the population of vectors was evolved using CR, F, and K.

The DE technique works successfully where other techniques fail (Synchronous and Fourier methods). This is the case when a noisy, wide bandwidth and/or closed fringe pattern is demodulated. Regularization techniques can be used in these cases but DE technique has the advantage that the cost function does not depend upon the existence of derivatives and restrictive requirements of continuity (gradient descent methods). Since the DE works with a population of possible solutions instead of a single solution, it avoids falling in a local optimum. Additionally, no filters and no thresholding operators were required, in contrast with the fringe-follower regularized phase tracker technique.

The DE has the advantage that if the user knows prior knowledge of the object shape, then a better suited fitting parametric function can be used instead of a general polynomial function. Additionally, due to the fact that the DE technique gets the parameters of the fitting function, it can be used to interpolate sub-pixel values and to increase the original phase resolution or interpolate where fringes do not exist or are not valid. A drawback is the selection of the optimal initial DE parameters (such as population size, F, K) that can increase the convergence speed.

Acknowledgements. We wish to thank the Centro de Investigación en Computación of the I.P.N. by the support to accomplish this project as well as the Centro de Investigaciones en Optica during the image recollections and tests. J. Vielma thanks CONACYT by the scholarship received to complete his doctoral studies. H. Sossa thanks the SIP-IPN under grant 20091421 for the support. H. Sossa also thanks CINEVESTAV-GDL for the support to do a sabbatical stay from December 1, 2009 to May 31, 2010. Authors thank the European Union, the European Commission and CONACYT for the economical support. This paper has been prepared by economical support of the European Commission under grant FONCICYT 93829. The content of this paper is an exclusive responsibility of the CIC-IPN and it cannot be considered that it reflects the position of the European Union. Finally, authors thank the reviewers for their comments for the improvement of this paper.

References

- [1] Martín, F. et al.; New advances in Automatic Reading of VLP's, Proc. SPC-2000 (IASTED), Marbella, España, 2000, 126-131.
- [2] Malacara, D., Servin, M., Malacara, Z.: Interferogram Analysis for Optical Testing, Marcel Dekker, New York, 1998.
- [3] Malacara, D.: Optical Shop Testing, Wiley, New York, 1992.

- [4] Creath, K. in: E. Wolf (Ed.), *Progress in Optics*, vol. 26, Elsevier, Amsterdam, 1988, p. 350.
- [5] Creath, K. in: D. Robinson, G.T. Reid (Eds.), *Interferogram Analysis*, IOP Publishing, London, 1993, p. 94.
- [6] Takeda, M., Ina, H., Kobayashi, S.: Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry, *Journal of Optical Soc. of America*, Vol. 72, 1981, pp. 156-16.
- [7] Su, X., Chen, W.: Fourier transform profilometry: a review, *Optics and Lasers in Engineering*, Vol. 35, Issue 5, May 2001, pp. 263-284.
- [8] Womack, K.H.: Interferometric phase measurement using spatial synchronous detection, *Opt. Eng.*, Vol. 23, 1984, pp. 391-395.
- [9] Servin, M., Rodriguez-Vera, R.: Two dimensional phase locked loop demodulation of interferograms, *Journal of Modern Opt.*, Vol. 40, 1993a, pp. 2087-2094.
- [10] Ghiglia, D.C., Romero, L.A.: Robust two-dimensional weighted and unweighted phase unwrapping that uses fast transforms and iterative methods, *J. Opt. Soc. Am. A*, Vol. 11, 1994, pp 107-117.
- [11] Su, X., Xue, L.: Phase unwrapping algorithm based on fringe frequency analysis in Fourier-transform profilometry, *Opt. Eng.* 40, 2001, pp 637-643.
- [12] Servin, M., Cuevas, F.J., Malacara, D., Marroquin, J.L., Rodriguez-Vera, R.: Phase unwrapping through demodulation by use of the regularized phase-tracking technique, *Appl. Optics*, Vol. 38, No. 10, 1999, pp. 1934-1941.
- [13] Servin, M., Marroquin, J.L., Cuevas, F.J.: Demodulation of a single interferogram by use a two-dimensional regularized phase-tracking technique, *Appl. Opt.* Vol. 36, 1997, pp. 4540-4548.
- [14] Villa, J., Servin, M.: Robust profilometer for the measurement of 3-D object shapes based on a regularized phase tracker, *Opt. Lasers Eng.* Vol. 31, 1999, pp. 279-288.
- [15] Quiroga, J.A., Gonzalez-Cano, A.: With a Regularized Phase-Tracking Technique, *Applied Optics*, Vol. 39, Issue 17, 2000, pp. 2931-2940.
- [16] Cuevas, F.J., Servin, M., Stavroudis, O.N., Rodriguez-Vera, R.: Multi-Layer neural network applied to phase and depth recovery from fringe patterns, *Opt. Comm.*, Vol. 181, 2000, pp. 239-259.
- [17] Cuevas, F.J., Servin, M., Rodriguez-Vera, R.: Depth object recovery using radial Basis Functions, *Opt. Comm.*, Vol. 163, 1999, p.270.
- [18] Servin, M., Marroquin, J.L., Cuevas, F.J.: *J. Opt. Soc. Am. A* 18 (2001) 689.
- [19] Price, K.V., R.M. Storn, and J.A. Lampinen, *Differential Evolution - A Practical Approach to Global Optimization*. 2005: Springer. 538.
- [20] Storn, R. and K. Price, *Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces*. *Journal of Global Optimization*, 1997. 11: p. 341-359..
- [21] Li, X. *Efficient Differential Evolution using Speciation for Multimodal Function Optimization*. in *Conference on Genetic and Evolutionary Computation*. 2005. Washington DC.
- [22] Robic, T. and B. Filipic, *DEMO: Differential Evolution for Multiobjective*. 2005, Jozef Stefan Institute: Slovenia. p. 520-533.
- [23] Roger, L.S., M.S. Tan, and G.P. Rangaiah, *Global Optimization of Benchmark and Phase Equilibrium Problems Using Differential Evolution*. 2006, National University of Singapore: Singapore.
- [24] Gamperle, R., S.D. Müller, and P. Koumoutsakos, *A Parameter Study for Differential Evolution*. 2006, Swiss Federal Institute of Technology Zürich: Zürich.

Enhancing the Diagnosis Module in a Self-healing Architecture Supporting Web Service Applications

Francisco Moo-Mena, Fernando Curi-Quintal, Juan Garcilazo-Ortiz, Luis Basto-Díaz, and Roberto Koh-Dzul

Universidad Autónoma de Yucatán, Facultad de Matemáticas,
Periférico Norte-Tablaje 13615, Mérida, Yucatán, Mexico
{mmena,cquintal,gortiz,luis.basto}@uady.mx, jose.koh@format.uady.mx

Abstract. A Self-healing infrastructure allows to observe the behavior of a system, determine its health status, and apply measures to restore the correct state of the application. In recent years our work has focused on the design and implementation of Self-healing architectures, which support applications based on Web services (WS). A previously developed architecture adopted a centralized approach regarding to the topology and control components. As a diagnosis technique used a statistical method based on box-plot diagrams. In this paper we present the modifications made to the diagnosis module of that Self healing architecture. The proposed change gets an architecture with distributed control components. The diagnosis also adds the use of ontologies and inference rules that contribute to improving awareness about system health. The results obtained by applying new Self-healing architecture to a distributed digital library application show a trend towards more precise diagnosis.

1 Introduction

The development of technology in recent decades has had a major impact on human lifestyle, creating a dependence on systems that handle storage, processing and management of meaningful information in a diversity of areas, like academic, industrial and services.

To meet the current needs, systems are more complex and their own maintenance and repair mechanisms are very hard to implement. It is a requirement their permanent availability, reliability and safety in heterogeneous contexts, which are susceptible to failures or malicious attacks.

The fault tolerance strategies and problem solving by managers are no longer sufficient in highly dynamic environments. New strategies are needed to have more available and efficient services [1].

Claiming to offer a solution to that kind of problems, IBM presented, in 2001, the Autonomic Computing initiative [2], inspired by the autonomous function of central nervous system of animals, adopting four perspectives:

1. *Self-configuring*. Systems adapting in response to their environment changes, based on high-level policies.
2. *Self-healing*. Ability to detect, diagnose and recover from errors.
3. *Self-optimizing*. Improves system performance by optimizing its operation and resource usage.
4. *Self-protecting*. Defense strategies are proposed to solve problems caused by attacks or failures that cannot be repaired by Self-healing.

Implementing these four perspectives results in systems that automate the maintenance, repair, optimization and protection tasks, requiring minimal human or other systems intervention.

An autonomous system is composed of autonomous elements that contain resources and offer services. The elements handle system's internal behavior and relationships with other elements according to established policies.

The Self-healing and Self-protecting approaches are interesting because they are focused in maintaining the permanent availability, accessibility and integrity of the system, by preventing and repairing faults.

However, from another point of view, Self-healing perspective are more important because a system may have the ability to auto-configure, self-optimize and self-protection, but if it cannot recover from failures is likely to stop working.

Some Self-healing approaches propose a process that helps to detect and recover system's faults. Their stages are [3], [4]:

1. *Monitoring*. Consists in the monitoring and recording of information about the system health status.
2. *Diagnosis*. Analyzes and evaluates the information gathered in the monitoring stage and determines whether the performance level is correct or not.
3. *Recovery*. Strategies are implemented to help correct problems found and recover the proper system performance.

This paper adopts an approach focused on the architecture of the application, applying interceptor techniques for monitoring, Quality of Service (QoS) analysis for the diagnosis and redundancy of components for recovery.

The aim of our project is to obtain an architecture based on WS, which allows the creation of a distributed system where the interaction between WS has Self-healing capabilities, ensuring the quality of communication among them.

We use a distributed digital library application in order to test our proposal. In this application each Web Service hosts information corresponding to an area of knowledge library [5], pretending create a distribution of content in different WS, ensuring each has access to the contents of the other. The reason for the distribution of content is to avoid a central point of failure, and improve their performance using the Self-healing scheme.

This work is based on the architecture and Self-healing techniques presented in [5], [6] and [7] in order to make a new proposal by adding a distribution perspective, and a better diagnosis process.

The rest of the paper is organized as follows: in Section 2 an overview of related work regarding Self-healing systems is presented. Section 3 introduces

the Self-healing architecture previously developed. Section 4 describes in general terms the new Self-healing architecture with emphasis on changes carried out to the diagnosis module. Section 5 shows experimental results. And Section 6 describes conclusion and ongoing work.

2 Related Work

In recent years, there have been different perspectives on Autonomic Computing concepts, leading to propose new initiatives like Self-protecting, Self-knowledge, Self-diagnosis, Self-destruction and Self-adjustment, according to emerging needs, fitting the concept of autonomy [3].

For the implementation of Self-healing systems the main problems encountered are designing the mechanisms for fault detection and diagnosis, and recovery strategies. Solutions have been proposed for specific cases that can be applied to other approaches. Hardware-based autonomous designs find their counterpart in software systems. Some projects related to Autonomic Computing that propose Self-healing approaches are:

2.1 Bio-Net

Led by the National Science Foundation, Bio-Networking Architecture [8] is a paradigm and a middleware for the design and implementation of scalable, adaptable and available network applications. It is based on principles and mechanisms used by biological systems to adapt to changing environmental conditions, like colonies of ants and bees. Abstracts the biological model of cooperation and autonomy in a system composed of autonomous agents, cyber-entities (mobile autonomous agents) and Bio-Networking Architecture platforms (execution environments and support services for the cyber-entities). The autonomous interaction and cooperation between the components results in a stable and adaptable survival system.

2.2 HYDRA

Hydra [9] is a middleware belonging to Hydra EU project, which allows developers to incorporate heterogeneous devices offering WS interfaces for administration.

Hydra incorporates mechanisms for Service Discovery, Semantic Model Driven Architecture, P2P communication and Diagnosis.

Hydra adopts the "awareness context" concept and presents an OWL ontology and SWRL rules based on the Self-management approach, particularly in Self-diagnosis. In Hydra, each component has an awareness of their own state, knowing and defining the optimal values for parameters benchmarked to run smoothly.

SWLR rules and ontologies provide a Knowledge Base where information about the current state of the middleware is represented. If a fault occurs the

possible solutions are within the same ontology OWL. The inference about the information stored during the monitoring helps the diagnosis process to make intelligent decisions regarding the recovery Self-healing tasks.

2.3 CODA

CODA (Complex Organic Distributed Architecture) [10] represents a new generation of decision-making system that includes a means monitoring and controlling objectives to allow the enterprise to evolve with certain degree of autonomy.

The architecture includes concepts and principles of Self-organization, Self-regulation toward an intelligent architecture. The main challenge is to achieve a distributed object-oriented reference architecture with support for reconfigurable mobile networks.

2.4 Discussion

Our work has focused on creating platforms and application models using the Self-healing perspective.

Previous works have presented concepts such as adaptation of the application's components to new environments, the "context awareness" for diagnosis (using ontological models), distribution of components into an intelligent architecture. Other works have developed concepts that focus on developing applications based on WS, applying Self-healing principles, for example, WS-DIAMOND [11].

Our goal is to develop a new perspective for the creation of WS-based applications, in which the cooperative interaction is guaranteed by applying Self-healing principles.

We have developed an ontological-statistical model for performance analysis, approaching to the "context awareness" based on QoS, which considers the reconfiguration of the techniques used for fault detection in dynamic environments.

Our case study consists in to implement a distributed WS-based digital library application in which each component is responsible for storage and information management of an area of knowledge.

3 Previous Self-healing Architecture

Nowadays, systems implementation with Self-healing properties has reached a considerable importance. Among the diverse approaches covered, the main problem worked is the components complexity. It is no longer enough to have fault tolerance mechanisms, additional strategies are needed to ensure an effective recovery. This paper presents improvements made to an architecture defined in [5].

The previous architecture is centralized and consists in a Web Service Consumer, a Web Service Provider and a Self-healing Core, where databases and Monitoring, Diagnosing and Recovering modules are located.

The architecture was implemented with the Java Web Service technology. Apache Axis2 is the Web Service engine, running the consumer and the provider. The Self-healing Core is a component whose interface with WS is through Java RMI.

3.1 Monitoring Module

Every transaction between WS consumer and WS provider is monitored using Apache Axis2 Handlers, making the role of interceptors, which are responsible to collect and save timestamps defined as follows:

- T1: Service Request's start time
- T2: Service Request's end time.
- T3: Service Response's start time.
- T4: Service Response's end time.

The lack of any of timestamps indicates a corrupted performance in the provider.

When the transaction is complete, the recorded timestamps are used to calculate QoS parameters of time, defined as follows:

- QoS1: $T3 - T2$: Computation time
- QoS2: $T2 - T1$: Requesting time.
- QoS3: $T1 - T3$: Responding time.
- QoS4: $(T1 - T1) - QoS1$: Communication time.

The parameters are retrieved from the PaRe database table after a certain number of transactions, to perform diagnosis operations.

3.2 Diagnosis Module

The diagnosis stage implements the QoS analysis strategy, using a statistical model to compare the set of values obtained from monitoring with a reference model defined by analyzing the performance of the architecture. The result let to determine whether or not any recovery action is required.

The statistical model, exposed in [6], is based in the box-plot method. The Interquartile Range (IQR) and Right Outer Fence ($Q3 + 3.0 * IQR$) were defined using samples of parameters from a first test. The portion of outlier values is observed, according to the measures established by each WS provider stored in the Service Level Parameter (SLP) table. The SLP table contains the measure known as "Left Outer Fence" for the QoS parameters of each WS with which it has interacted. During the first diagnosis process, SLP contains values that were calculated in ideal conditions and would represent good performance.

The model determines the percentage of outliers values (greater than Right Outer Fence) that are in a set of data collected. If the percentage is between 0% and 5% means that the health status is correct, between 5% and 10% indicates that the WS provider is degrading, and a second WS provider is requested to divide the work (Duplication). If the percentage is greater than 10% then the WS provider is degraded and other WS provider is necessary (Substitution).

3.3 Recovery Module

Recovery strategies are based on the system's component redundancy. The duplication and substitution of the WS provider allow that requests made by the WS consumer are met, avoiding results loss. Recovery depends on the availability of redundant WS.

Inside of the Self-healing Core is the WSTA database (Web Service Table Access) that stores information about the WS providers available. The records shall indicate the endpoint, status (online, offline, active, inactive), a description and operations offered by the WS. Records in the database are defined by an external agent, such as an administrator. The architecture does not add or delete information.

For the execution of recovery actions, the records in the WSTA are modified. Duplicating a WS is reflected as the activation of more than one provider (change of the *active* attribute). The WS substitution consists in disable the current provider (*active=false*, *offline=true*) and activate another available WS (*active=true*).

3.4 Overview Operation

The operation of the architecture is simple, keeping the Self-healing principles. The operation begins with a WS consumer request. To send it, a WSTA database query is performed requesting the most appropriate WS provider available (the database query includes the action and description required). When the WS consumer receives the endpoint to the WS provider, sends the request, initiating a monitored transaction; in case of failure, a recovery action is performed. Each transaction between the provider and the consumer follows the same process, whether they are equal.

After a certain number of transactions, the diagnosis module executes to detect and correct possible faults.

3.5 Critical Analysis of the Previous Architecture

The previous architecture has limitations with regard to its structure and not allowing the dynamism of WS.

The main limitation of the architecture is the centralized and rigid approach, only allowing a WS consumer interacts with one or more WS providers. Besides setting out a critical point, the Self-healing Core, where a fault may represent the complete loss of functionality, since the absence of recovery mechanisms on this component.

But the key point lies in the diagnosis module, where the QoS analysis is limited to observing time parameters of transactions. The dispersion measures applied do not consider variations in the interactions time intervals. In the diagnosis module, the WSTA table hasn't the ability to upgrade itself according to actual active providers, it depends on an update by the administrator.

It is necessary to accomplish the architecture distribution, improve the diagnosis process and establish adaptive mechanisms for the statistical model.

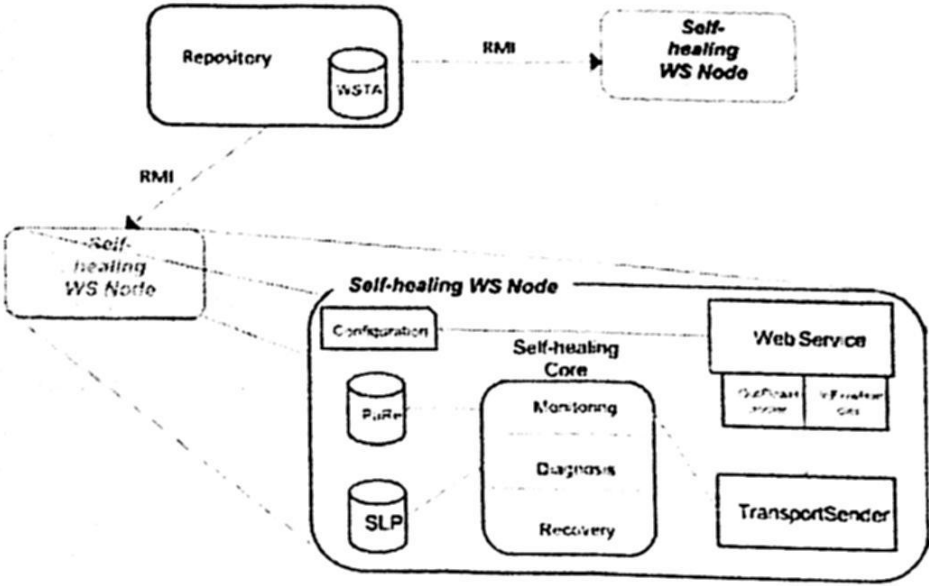


Fig. 1. Distributed Self-healing architecture.

4 Enhanced Distributed Self-healing Architecture

In our work we have made changes in the former architecture, improving Self-healing features, which were restricting the WS performance. This section presents the changes in the architecture's components. According to the paper's title, the major changes are in the diagnosis module.

The components distribution adapts the architecture to the dynamic interaction of the WS. WS can interact with each other according to their needs at any given time.

According to the project's objective, a digital library with distributed contents is obtained when each WS plays either provider or consumer roles, even both simultaneously.

Then each WS monitors the performance of its providers, assesses and determines the need to implement a recovery action, either changing provider or sending a request to more of them. The Self-healing Core, which was an isolated component, now is part of the Web Service. Each WS contains its own Self-healing components (Fig. 1).

Distributing components avoids critical points of failure that could have major consequences. The failure of a WS or a stack of Self-healing components has no effect beyond that WS. A serious flaw could require the full restoration of a WS, at most.

4.1 Repository Module

Actually, WSTA database is an independent component with a similar structure to the previous Self-healing Core. Its function is to register the WS availability, besides being the architecture's time server for consistency of timestamps.

Registering WS corresponds to each WS that integrates itself into the architecture. The changes applied to the records, as part of recovery actions, are visible to all other WS in the architecture.

4.2 Monitoring Module

The monitoring process has no major changes. The records and parameters are stored in local databases in every WS and are divided by each interacting WS provider.

The WS playing the consumer role in any interaction performs the monitoring.

4.3 Diagnosis Module

This module, in the enhanced architecture, will continue applying the defined statistical model, but is now complemented by an ontological model responsible for making inferences about the performance of the WS provider and setting the conditions to require a repairing action.

Adaptation of the reference measurements. The analysis of the observed parameters returns the percentage of outliers in the set according to the Right Outer Fence defined in the model.

A portion less than 5% indicates that the provider's performance is good, although there may be signs of increased time intervals. Values greater than 2.5% could indicate that the parameters are changing, perhaps due to changes in the underlying networking system.

Ignoring possible variations, the following diagnosis operations may invoke recovery actions, even if everything is working properly. Variations in parameters must be considered since the Right Outer Fence defined for a first test does not apply to sets of parameters in all contexts. Adapting the model's descriptive measures is required.

In the new diagnosis implementation, if the percentage of outliers parameters is between 3% and 5%, a necessary measure adaptation action is established. The changes are stored in the SLP table for further use. The box-plot method is executed using the set of values for each QoS parameter, determining a new Right Outer Fence's value giving a new diagnosis' perception.

If a set of values of a QoS parameter requires a measures adaptation, the same procedure for the three remaining parameters is performed.

The variation in one parameter may involve variations in the other, so if a parameter requires their descriptive measure redefinition, it is likely that someone else also makes that request. Executing adaptation strategies for all parameters prevents possible future recurrences.

The first execution of the diagnosis process uses the descriptive measurements defined in the previous architecture and according to the results may or may not request a re-calculation (adaptation). The adjustment is incremental in the

$s \in \text{status} \cdot \text{HasParameter}(s, \text{qos}) \cdot \text{GreaterThan}(\text{qos}, \text{qosdip}) \cdot$ $\text{LessThan}(\text{qos}, \text{qossub})$ $\Rightarrow \text{HasDegradation}(s, \text{moderate})$ $s \in \text{status} \cdot \text{HasDegradation}(s, \text{moderate})$ $\Rightarrow \text{NeedRecovery}(s, \text{duplication})$
--

Fig. 2. Rules in diagnosis module.

Left Outer Fence value; no change in this measure indicates system's operation stability.

When the implementation of recovery actions is determined, any kind of adaptation does not take place, since degradation might be ignored.

Finally, the results of the statistical model are sent to the ontological model to proceed with the stage of diagnosis.

Diagnosis based on ontologies. Diagnosis based on the statistical method may be considered incomplete. WS performance is reflected in several factors.

The QoS parameters such as availability, accessibility, integrity, performance, reliability, control and safety [12] may also be appropriate indicators of the WS provider performance.

To get an insight into these parameters, the WS must be aware of their own state and act to prevent or recover from failures or degradations in its performance.

Ontologies are an alternative to model the environment and the state of the WS, showing capacity to make inferences, using logical rules, to determine a good or bad performance.

In a previous work [7] we defined an ontological model that includes QoS parameters and rules to infer the current state of architecture. In the present work, the Self-healing architecture's diagnosis module is attached to a rules engine that is responsible for analyzing the data collected by monitoring, using the ontological model. So far, the parameters analyzed were the time intervals already defined.

The change resides in the fact that results are no longer mere assertions, now consist of logical conclusions by applying first-order logic sentences as shown in Fig. 2.

For implementation, we used Jess rule engine [13], which use OWL-DL ontologies via the Protégé and its complement JessTab. With those tools we build a package that provides methods to send and receive data from the ontological model. This package is accessible from the methods implemented in architecture components.

When the ontological model receives the necessary parameters from the statistical model, those are represented in the ontology and the logical rules compare the current performance with a reference model for diagnosis

4.4 Recovery Module

The Recovery Module implements strategies to recover an appropriate of performance when the diagnosis module may determine necessary. Currently, each WS applies recovery strategies from its recovery module.

The strategies are techniques of component redundancy and consist in changes in WSTA information indicating the activation or inactivation of WS providers in the architecture.

WS modifications over the WSTA database do affect the operation of everyone. If a provider is set to offline, will not receive requests from customers, however, it can continue with the role of consumer.

5 Results

The main result to present is the adaptation of the statistical model's measures.

In the test showed in Fig. 3, we obtained a set of values of QoS1 parameter during the interaction between WS "botany" and WS "math". If this set is analyzed using the default Right Outer Fence, defined in an earlier trial (365.5), the percentage of outlier values is 3.16% of the total (40) value obtained from the previous architecture (See Fig. 4).

Implementing the adaptation strategy, in the enhanced diagnosis module, gets a new $IQR = 142.5$ and redefines the Right Outer Fence in 585.5. In this case, the percentage of outliers values is 0%, preventing any unnecessary recovery action.

The value 0% is transferred to the ontology, inferring a good performance.

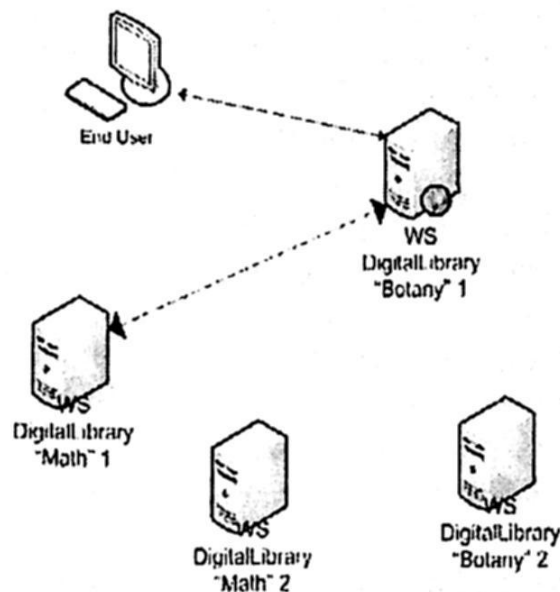


Fig. 3. Test scenario.

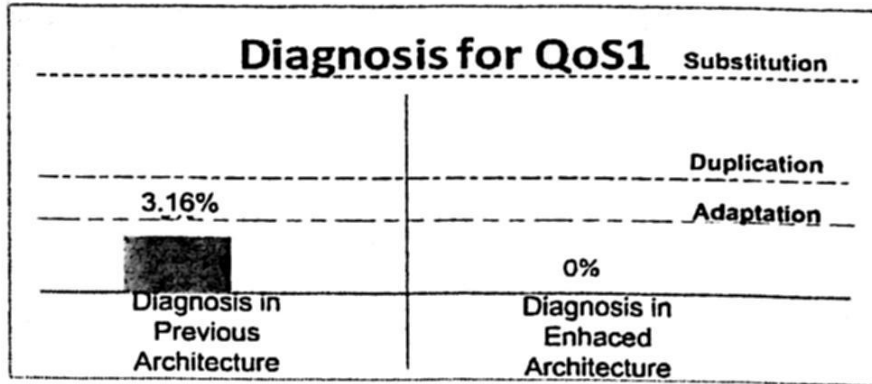


Fig. 4. Comparing diagnosis modules.

6 Conclusion

All changes applied to previous Self-healing architecture were not complicated to apply and contributed to obtain better results.

Distribution added to previous architecture allowed the possibility to distribute WS applications modules.

Quality of service reference measures updated constantly, allowed to architecture precision about the application healing, by providing the capacity to identify normal changes from real failures.

Current Self-healing architecture allows modifying constantly QoS measures, adapting to new conditions. However, adaptation is incremental as a future work that could be improving finding the best measures to be applied on each transaction.

An ontology module integrated into diagnosis, allowed other information about WS provider's performance giving better results. By now, only time parameters are considered, the addition of more parameters will be a future work.

7 Acknowledgments

This project is developed under financial support from PROMEP and UADY.

References

1. Shaw, S.: "Self-Healing": Softening Precision to Avoid Brittleness. Proceedings of the first workshop on Self-healing systems, pp. 111-114. ACM, New York, USA (2002)
2. Kephart, J., Chess, D.: The vision of autonomic computing. *Computer*, 1(36):41-50 (2003).
3. Halima, R., Drira, K., Jmaiel M.: A comparative study of self-healing architectures in distributed systems. *Rapport LAAS No 06568* (2006).
4. Ghosh, D., Sharman, R., Raghav, H., Upadhyaya S.: Self-healing systems - survey and synthesis. *Decision Support Systems* 42, pp. 2164-2185 (2007).

5. Moo-Mena, F., Garcilazo-Ortiz, J., Basto-Díaz, L., Curi-Quintal, F., Alonzo-Canul, F.: Defining a SelfHealing QoS based Infrastructure for Web Services Applications. In: IEEE 11th International Conference on Computational Science and Engineering, pp. 215-220. IEEE Press (2008).
6. Moo-Mena, F., Garcilazo-Ortiz, J., Basto-Díaz, L., Curi-Quintal, F., Medina-Peralta, S., Alonzo-Canul, F.: A Diagnosis Module Based on Statistic and QoS Techniques for Self-healing Architectures Supporting WS based Applications. In: IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE Press (2009).
7. Moo-Mena, F., Garcilazo-Ortiz, J., Basto-Díaz, L., Curi-Quintal, F., y Canul-Centeno, F.: Una ontología para el diagnóstico de la QoS en aplicaciones basadas en servicios Web (WS). In: XXII Congreso Nacional y VIII Congreso Internacional de Informática y Computación ANIEI (2009).
8. Wang, M., Suda, T.: The bio-networking architecture: A biologically inspired approach to the design of scalable, adaptive, and survivable/available network applications. Tech. Rep. 00-03, Department of Information and Computer Science, University of California, Irvine, California (2000).
9. Zhang, W., Hansen, K.: Towards Self-managed Pervasive Middleware using OWL/SWRL ontologies. In: Fifth International Workshop on Modeling and Reasoning in Context, pp. 1-12. (2008)
10. Ribeiro-Justo, G., Karran, T.: An Object-Oriented Organic Architecture for Next Generation Intelligent Reconfigurable Mobile Networks. In: Blair, Gordon, (ed.) DOA'01: 3rd International Symposium on Distributed Objects and Applications, pp. 31-40. IEEE Conference Proceedings . IEEE Computer Society, Las Alamitos, USA (2001).
11. Console, L., Fugini, M.: WS-Diamond: An Approach to Web Services, Diagnosability, Monitoring, and Diagnosis. tech. report 2007.57, Dept. Electronics and Information, Politecnico di Milano. (2007).
12. IBM: Understanding quality of service for Web services. <http://www.ibm.com/developerworks/library/ws-quality.html>
13. Ernest Friedman-Hill: Jess The Rule Engine for the Java™ Platform Version 7.1p2. Sandia National Laboratories. (2008).

Some Experiments on Grammatical Inference of English, Spanish and Pseudo-English*

David Pinto, Mireya Tovar, Sofía Paniagua, Beatriz Beltrán, Darnes Vilariño

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, Mexico

{*davideduardopinto,mireyatovar,sofiapaniagua,bbeltranmtz*}@gmail.com

Abstract. Grammatical Inference (GI) is one of the most promising areas of artificial intelligence, not only because its possible use in classification, but for its new items generation capability, since GI creates a grammar from which it is possible to generate new strings which in fact belong to the category modeled by that grammar. In this paper we present the results of inferred grammars for three human beings sublanguages written in "English", "pseudo-English" and, "Spanish" gathered to conform two different corpora: MATIS and DIHANA. We have used a grammatical inference method based on mixtures of bigrams to model those grammars, observing that this method allow enough generalization to accept in average the 60% of never seen strings. Besides, we have done an analysis of the original corpora, and our findings indicates that a redefinition of both, the training and test set must be done in order to obtain worst-case results, instead of average cases. Therefore, with the new corpora, the grammatical inference task comes to be much more difficult.

1 Introduction

Grammatical Inference is the process of learning of grammars and languages from data [6]. It is often referred to as automata induction, grammar induction, and automatic language acquisition. This area of the artificial intelligence deals with the deduction of a grammar from given examples (instances). It is also known as "inductive inference" and recently as "computational learning". Grammatical inference can also be defined as the process of learning an automaton from a set of string samples [3], since a grammar can be modeled by an automaton, without loss of generality. The classical approach to grammatical inference was first given by Gold [4], who introduced the notion of identification in the limit, which concerns with the behaviour of an inference algorithm on an infinite sequence of samples in the limit. We can formulate a formal description of that as follows. A complete presentation of a language L is an infinite sequence of ordered pairs $(w, l) \in \Sigma^* \times \{0, 1\}$, where $l = 1$ if $w \in L$ and 0 otherwise, and every string $w \in \Sigma^*$ appears at least once.

* This work has been partially supported by the CONACYT project #106625, as well as by the PROMEP/103.5/09/4213 grant.

If an inference method M is executed over a complete presentation, it will generate an infinite sequence of guesses g_1, g_2, g_3 , etc. M is said to identify L in the limit if there exists some number n such that all of the guesses g_i are the same for $i \geq n$, and g_n is equivalent to L . Further details concerning the Grammatical Inference (GI) field are well given in [9] and [8].

Unfortunately, obtaining an infinite number of samples is not viable and, therefore, we must investigate different ways of obtaining such language. In this paper we use a bigram-based mixture modelling to infer grammars for two different corpora. The mixture modelling is a well-known density estimation technique in supervised and unsupervised pattern classification [7]. Mixtures are flexible enough for finding a balance between the complexity of a model and the amount of training data available. The model complexity is often driven by varying the number of the mixture components, keeping the same parametric model (usually a very simple one) for all the components. Besides, the mixture maximum likelihood parameters estimation may be calculated by an Expectation-Maximization (EM) algorithm. This kind of algorithm is in fact the one we have used in our approach. Our experiments have made use of a bigram-based mixture to carry out the grammatical inference process. The system used was implemented using the fundamentals discussed in [2].

The next section describes the datasets used in our experiments. Section 3 presents the obtained results and, finally our conclusions are given.

2 Datasets

2.1 MATIS

This collection is a reduced version (Mini) from that denominated ATIS (Air Travel Information System), which was designed to measure the research advance in spoken language systems which includes both, speech and natural language components. For a better reference about ATIS see [5].

MATIS is made up by a *training* and a *test* set, both conformed by strings in natural language (English) referring to a partial part of a conversation from a user of an air travel information system. This corpus was preprocessed transliterating each natural language string in order to obtain a uniform language (pseudo-English) which can be better used in a computer system. MATIS was also preprocessed in order to contain "classes" which may help in the experiments carried out with this corpus; thus, the class $\langle \text{city} \rangle$ generalize instances like "New York", "Chicago", etc; whereas $\langle \text{airport} \rangle$ generalize instances like "JFK", "Newark", etc. However this preprocessing task has been applied with some mistakes, for instance, the next string was incorrectly rewritten: *SHOW ME ALL PRICE OF $\langle \text{day} \rangle$ CLASS FROM $\langle \text{city} \rangle$ TO $\langle \text{city} \rangle$* . Here we can see that the word "FIRST" was generalized by the class " $\langle \text{day} \rangle$ " which in fact is incorrect. Table 1 shows examples of the MATIS strings in natural language (NL) and uniform language (UL).

The original MATIS collection is made up by 2566 pairs of NL and UL strings; 2000 strings were selected for *training* and 566 for *test*. However, a simple

Kind	Strings
NL	- WHAT'S LOWEST ROUND-TRIP FARE FROM <city> TO <city>
NL	- DOES <airline> FLY FROM <city> TO <city>
NL	- WHAT'S EARLIEST FLIGHT FROM <city> TO <airport> THAT SERVE LUNCH
UL	- LIST CHEAPEST ROUND-TRIP FARES FROM <city> AND TO <city>
UL	- LIST <airline> FLIGHTS FROM <city> AND TO <city>
UL	- LIST EARLIEST FLIGHTS FROM <city> AND TO <airport> AND SERVING LUNCH

Table 1. MATIS string instances

analysis of the collection content has shown that it contains repeated strings. This feature is expected in real situations, but the worst-case analysis is the most common one used in the experiments. We have conformed a new version of MATIS with unrepeated strings which we have named URMATIS, and we will use the tags MATIS_NL and MATIS_UL for the natural language and uniform language versions of the original corpus, respectively; whereas we will use the tags URMATIS_NL and URMATIS_UL for the natural language and uniform language versions of the new preprocessed corpus, respectively. Table 2 shows the total number of strings for each corpora.

MATIS version	Training strings	Test strings
MATIS_UL	2000	566
URMATIS_UL	489	80
MATIS_NL	2000	566
URMATIS_NL	1534	360

Table 2. Different versions of MATIS

2.2 DIHANA

DIHANA is a corpus of Spanish spoken dialogs acquired and tagged in order to study and develop a robust dialog system for accessing information which contains spontaneous speech in a wide variety of environments. The DIHANA corpus content is about services, timetables and fares of the Spanish nation railtrain system. Thus, simulated conversations between a user and a wizard whom answer the user questions were recorded. DIHANA is made up by 6280 user interventions and 9133 system interventions. In the experiments carried out in this paper, we have used only the 4139 unrepeated user phrases, which we will further refer as SUBDIHANA. A better description of the original DIHANA corpus can be found in [1], and in the official website of the DIHANA project¹.

¹ <http://www.dihana.upv.es/>

In Table 3 we show examples of the DIHANA strings in their original language (Spanish) and a translation of each string, for a better understanding of this paper. As can be seen, this corpus is written in natural language and therefore it does not take into account possible classes implicit in the dialogs. Thus, we consider it a difficult collection for grammatical inference task.

Strings
quisiera saber los horarios de tren desde Valencia a Barcelona (I would like the train timetables from Valencia to Barcelona)
querria un tren desde Segovia hasta Badajoz (I would like to take a train from Segovia to Badajoz)
pues me gustaria salir manana temprano (I would like to leave tomorrow morning)
quiero el precio del viaje ida y vuelta (I want the roundtrip fare)
muy bien el lunes (monday is ok)

Table 3. DIHANA string instances

3 Experimental results

A preliminar analysis of both corpora, MATIS and DIHANA was done in order to evaluate the generalization quality of the inferred grammar by means of a bigram-based mixture method. Table 4 shows the number of test strings accepted and rejected by the inferred grammar for the training set.

Corpus	Accepted	Rejected
MATIS_UL	542 (95,76%)	24 (4,24%)
URMATIS_UL	57 (71,25%)	23 (28,75%)
MATIS_NL	379 (66,96%)	187 (33,04%)
URMATIS_NL	176 (48,89%)	184 (51,11%)
SUBDIHANA	398 (62,28%)	241 (37,72%)

Table 4. Evaluation of a grammar, inferred from the *training* set

The high degree of accepted strings over the original corpora of MATIS (MATIS_UL, MATIS_NL) is quite expected since there exist strings in the *test* set whose in fact were used to infer the grammar, i.e., those strings also exist in the training set.

3.1 v -fold cross-validation evaluation

The v -fold cross-validation technique allows to evaluate how well the experiment “performs” when is repeatedly cross-validated in different samples randomly drawn from the data. Consequently, the results obtained will not be casual through the use of a specific distribution of the data collection. In the experiments we conducted, we joined each training and test set and then we splitted the obtained data in v partitions from which 1 is taken as a test set and the other $v - 1$ partitions are used as a training set. This procedure is carried out v times and the average of the executions is presented in Table 5, with $v = 6$ and $v = 10$.

The execution over the MATIS corpora obtained a better performance with the uniform language strings than those in natural language. This fact is also expected, since the uniform language uses a set of keywords which defines a kind of reduced grammar compared with the grammar of the natural language. The bigram-based mixture model used allow enough generalization for accepting in average the 60% of never seen strings for both corpora: MATIS and DIHANA.

Corpus	Partitions	Accepted	Rejected
URMATIS_UL	6	71,16 (75,18%)	23,5 (24,82%)
URMATIS_UL	10	43,4 (76,41%)	13,4 (23,59%)
URMATIS_NL	6	178 (56,45%)	137,33 (43,55%)
URMATIS_NL	10	109,2 (57,72%)	80 (42,28%)
SUBDIHANA	6	415,83 (60,29%)	273,83 (39,71%)
SUBDIHANA	10	253,7 (61,30%)	160,1 (38,69%)

Table 5. v -fold cross-validation evaluation

4 Conclusions

The inference of natural language dialogs is one of the most difficult task that can be done in grammatical inference, since in this task there exist only positive samples and, the only way of obtaining the complete grammar is by using an infinite number of samples which in fact is impossible. Moreover, it is really difficult to determine which phrase really belongs or not to the language because in theory any phrase should be part of the human language.

We have carried out an analysis of a grammatical inference system performance which is based in the use of bigram mixtures. The experiments were executed over two corpora: MATIS and DIHANA with three different languages: “English”, “pseudo-English” and, “Spanish”. We observed that the algorithm is capable of generalize quite enough to accept around the 60% of never seen strings which in fact belong to the training set inferred grammar.

References

1. N. Alcácer, J. M. Benedí, F. Blat, R. Granell, C. D. Martínez, F. Torres: *Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus*, Specom 2005.
2. J. Civera, A. Juan: *Mixtures of IBM Model 2*, In Proc. of the 11th Annual Conf. of the European Assoc. for Machine Translation (EAMT 2006), pages 159-167, Oslo (Norway), jun 2006.
3. M. L. Forcada: *Neural Networks: Automata and Formal Models of Computation*, Universitat d'Alacant, Dept. Llenguatges i Sistemes Informàtics, E-03071 Alacant (Spain).
4. E.M. Gold: *Language identification in the limit*, Information and Control, 10:447-474, 1967.
5. C.T. Hemphill, J.J. Godfrey, & G.R. Doddington, *The ATIS Spoken Language Systems Pilot Corpus*, DARPA Speech and Natural Language Workshop, Hidden Valley PA, June 1990.
6. V. Honavar and G. Slutzki (Ed), *Grammatical Inference*, Berlin: Springer-Verlag, 1998.
7. A. K. Jain, et al.: *Statistical Pattern Recognition: A Review*. IEEE Trans. on PAMI 22: 437, 2000.
8. L Pitt: *Inductive Inference, DFAs and Computational Complexity*, In J Siekmann (editor), *Proceedings of the International Workshop AII 89*, Lecture Notes in Artificial Intelligence 397, pages 18-44, Springer-Verlag, 1989.
9. Y. Sakakibara: *Recent advances of grammatical inference*, Theoretical Computer Science, 185:15-45, 1997.

A Grammatical Inference System based on Genetic Algorithms*

David Pinto, Beatriz Beltrán, Sofía Paniagua, Mireya Tovar, Darnes Vilariño

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, Mexico
{*davideduardopinto, bbeltranmtz, sofiapaniagua, mireyatovar*}@gmail.com

Abstract. Automatic learning of a grammar from a set of samples is matter of the grammatical inference field. Different approaches have been presented in literature for this challenging task. In this paper, we present the results obtained after the implementation of a genetic algorithm applied to the grammatical inference research area. Our system generates a main automaton which models the grammar we would infer and, thereafter, we obtain a set of positive and negative samples from this automaton for training a learning system. A set of automata are then generated as population and, after a number of generations, a similar grammar to the one represented by the main automaton is obtained. Several runs have been executed in order to calculate the mean performance of the presented approach. We have observed that the application of a genetic algorithm can highly improve the results, even if the initial and randomly-generated automata set has initial low scores.

1 Introduction

Grammatical inference can be defined as the process of learning an automaton from a set of string samples [1]. The classical approach to grammatical inference was first given by Gold [4], who introduced the notion of identification in the limit, which concerns is about the behaviour of an inference algorithm on an infinite sequence of samples in the limit. We may then give a formal description of this process as follows:

A complete presentation of a language L is an infinite sequence of ordered pairs $(w, l) \in \Sigma^* \times \{0, 1\}$, where $l = 1$ if $w \in L$ and 0 otherwise, and every string $w \in \Sigma^*$ appears at least once. If an inference method M is executed over a complete presentation, it will generate an infinite sequence of guesses g_1, g_2, g_3 , etc. M is said to identify L in the limit if there exists some number n such that all of the guesses g_i are the same for $i \geq n$, and g_n is equivalent to L . Further details concerning the Grammatical Inference (GI) field are completely described in [2] and [3].

Unfortunately, obtaining an infinite number of samples is not viable and, therefore, we must investigate different ways of obtaining such language. In this

* This work has been partially supported by the CONACYT project #106625, as well as by the PROMEP/103.5/09/4213 grant.

research work, we have used the genetic algorithms field in order to experiment with grammatical inference. Genetic algorithms, introduced by Holland in [6] at the middle of the 70s, are based on the mimic of the natural selection. Population, selection, crossover and mutation are topics well known in this field. The approach presented in this paper comprises the use of a set of random automata as population and, thereafter, to apply a genetic algorithm process in order to obtain, after a number of generations, an automaton similar to the language L . The language is represented without loss of generality by an automaton from which a set of positive and negative samples are obtained.

The rest of this paper is structured as follows. In Section 2 we describe the general approach of this research work. Section 3 describes each component of the presented GI-based system into detail. In Section 4, a set of evaluations carried out are presented and discussed. Finally, in Section 5 the conclusions are given.

2 Inference through different small grammars

The aim of this research work is to infer one grammar through a set of positive and negative samples. The usual approach consists in constructing a fixed grammar that may recognize all the training samples and, thereafter, to generalize this grammar in order to accept more samples which usually are not represented in the training set of samples. The contribution of this paper consists in the use of not only one but a set of grammars in order to generalize the expected grammar that will accept the samples given in the training set and other ones which are considered to be part of the analysed language. In other words, we would like to investigate whether or not, a grammar may be described by a set of grammars with a small number of components. This approach would be useful since, the traditional approaches use huge automata and, therefore, they are expensive in terms of computational time. The representation of the presented approach is given in Figure 1. Without loss of generality, in this paper we will refer to grammars and automata as the same concept, since any grammar may be represented by means of some automaton.

The set of small automata are considered to be calculated on the basis of a genetic algorithm which evaluates in each iteration the fitness of the current population of automata.

3 Description of the implemented system

In order to correctly tune and analyse datasets, an experimental system requires different tools to allow reproduce, once and once, the experiments carried out. The system we have built is made up by a set of modules which are described into detail in the following subsections.

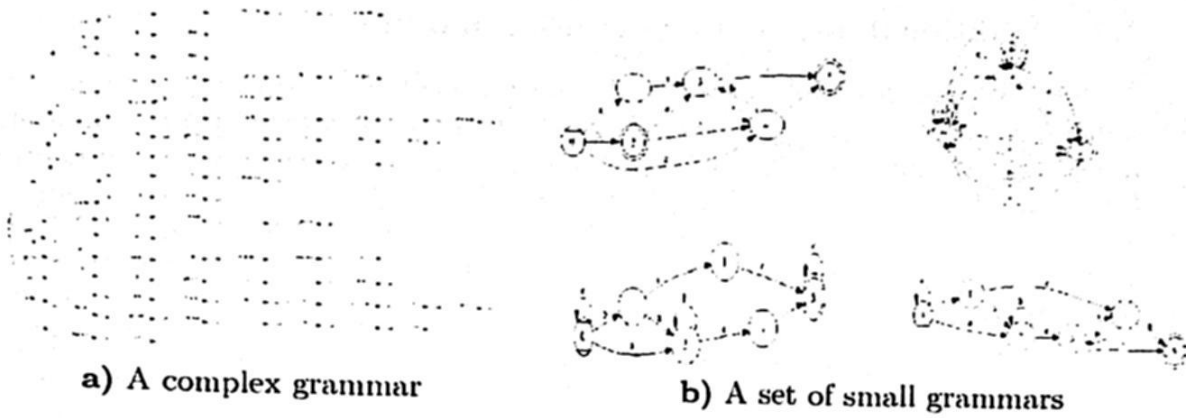


Fig. 1. Representing one complex grammar through a set of small grammars

3.1 Main automaton generation module

This module was implemented for constructing a main automaton¹. This automaton will be used to generate a set of positive and negative samples from which we would infer a grammar. Although only the main automaton may be used to evaluate how well the new grammar was inferred, another function of evaluation was implemented (see Subsection 3.4). Details corresponding to this module are shown as follows. We have included the parameter *NUM_STATES*, which defines the number of states of the main automaton. The *ARCS_THRESHOLD* parameter is a value between 0 and 1 which allows to decide whether an arc from a source node to a target one will be defined or not. This procedure is carried out for each node (state), with every symbol (*NUM_SYMBOLS*). Finally, the *FINALS_THRESHOLD* parameter will be used for determining, for each state, their possibility to be final (end state). For convenience, when an automatic generation of the automaton is performed, we have defined letters to identify every arc (symbol) in the graph ('a', 'b', 'c', etc).

3.2 Samples generation module

The generation of positive and negative samples is carried out in this module. The parameters *NUM_POS_SAMPLES* and *NUM_NEG_SAMPLES* are used for determining the number of positive and negative strings to generate, respectively. We established the main automaton in "generation mode" in order to generate positive samples, that is, we start from the initial state and, thereafter, we select the next state by means of a equiprobable distribution of all leaving arcs from the source state. Every time a final state is reached, it is decided with probability *END_THRESHOLD* if the current string will be returned or not.

In the case of the negative samples, we have decided to generate random strings and, thereafter, to evaluate its membership to the grammar defined by the main automaton.

¹ The presented approach uses only finite and non-deterministic automata.

3.3 Additional automata generation module

The aim of this research work was to obtain a set of automata capable to figure out, as a whole, of the grammar defined by the main automaton. Therefore, we have implemented this module which carries out the same process defined in 3.2, but in this particular case, for a *NUM_AUTOMATA* number of automata.

3.4 Evaluation module

This module is decomposed into two parts: 1) the assessment of each automaton through the training samples and, 2) the assessment of a test set. The next subsections describe each component into detail.

Evaluation of training samples Given two sets of samples:

$$PS = \{P_1, P_2, \dots, P_{NUM_POS_SAMPLES}\}$$

and

$$NS = \{N_1, N_2, \dots, N_{NUM_NEG_SAMPLES}\}$$

, the evaluation procedure of one automaton A is carried out in the following manner. For each sample $S \in PS$, if S is accepted by A , then the $Score(A)$ is incremented, otherwise it is decremented. The treatment of negative samples is carried out in an inverse manner, i.e., for each sample $S \in NS$, if S is rejected by A , then the $Score(A)$ is incremented, otherwise it is decremented.

Evaluation of the test set After the evaluation of each automaton with the training samples, every automaton obtains a particular score. Logically, the automaton with the highest positive score will be the most similar one to the main automaton, as well as the complemented version of the automaton with the lowest negative score. Automata with score close to zero will accept or reject almost all of the training samples and, therefore, they do not have high influence in the evaluation results. Thus, we can use the score of each automaton to evaluate a test set.

Let us denote as *NUM_TEST_SAMPLES* the complete number of positive and negative samples used in the test procedure. Let $Score(A_j)$ be the score of the A_j -automaton and, t_{ij} a tag which indicates whether the A_j -automaton accepts or rejects the i -th test sample ($1 \leq i \leq NUM_TEST_SAMPLES$). t_{ij} will be equal to 1 if the A_j -automaton accepts the i -th test sample and, equal to -1 otherwise. Thus, by using the above remarks, the evaluation procedure may be done as presented in Equations (1), (2), and (3).

$$d_i = \sum_{j=1}^{NUM_AUTOMATA} t_{ij} \cdot Score(A_j) \quad (1)$$

$$c_i = \begin{cases} 1, & \text{if } d_i > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$accuracy = \frac{\sum_{i=1}^{NUM_TEST_SAMPLES} c_i}{NUM_TEST_SAMPLES} \quad (3)$$

In the presented GI-based system, the evaluation procedure may be done either by using the complete set of automata or, just with the best automaton. In general, an automata subset may be chosen by using a confidence threshold, which will be based on the maximum automaton score (*MAXIMUM_SCORE*). It is obvious that the maximum score that may be achieved by any automaton of a population will be equal to *NUM_TEST_SAMPLES*. The score of each automaton with respect to the maximum score may be obtained in the following manner:

$$Relative_score(A) = \frac{Score(A)}{MAXIMUM_SCORE}$$

Therefore, given a threshold T , an automaton A belongs to the evaluated set if $Relative_score(A) \geq T$. In the experimental results shown in Section 4, we have used the accuracy formula presented in Equation (3) for determining the quality of the system.

3.5 Genetic algorithm module

This module draws the typical approach of a genetic algorithm. An overview of this implementation can be seen as follows:

- Generate initial population
- For N number of generations do
 - Select two parents: p_1, p_2 (SELECTION)
 - Crossover p_1 and p_2 for obtaining two offsprings: o_1, o_2 (CROSSOVER)
 - Evaluate fitness of o_1 and o_2 (FITNESS EVALUATION)
 - * If any of the offsprings outperforms the worst population citizen, then the offspring substitutes the found citizen
 - If necessary, apply MUTATION
- End for

The SELECTION process was applied by using a uniform distribution among all the population members. The CROSSOVER was done by splitting each parent by two. Given m states, we took the first $m/2$ states of p_1 in order to conform the first part of o_1 , whereas the next $m/2$ states make up the second part of o_2 . We did a similar process with the other parent. A graphical view of this process can be seen in Figure 2.

The fitness evaluation of each offspring was carried out by using the evaluation function described in 3.4. Finally, the MUTATION phase was done only over a certain number of times (*MAX_VALUE_MUTATION*). We incremented a counter (*mutation_counter*) each iteration in which all the offsprings did not outperformed the population.

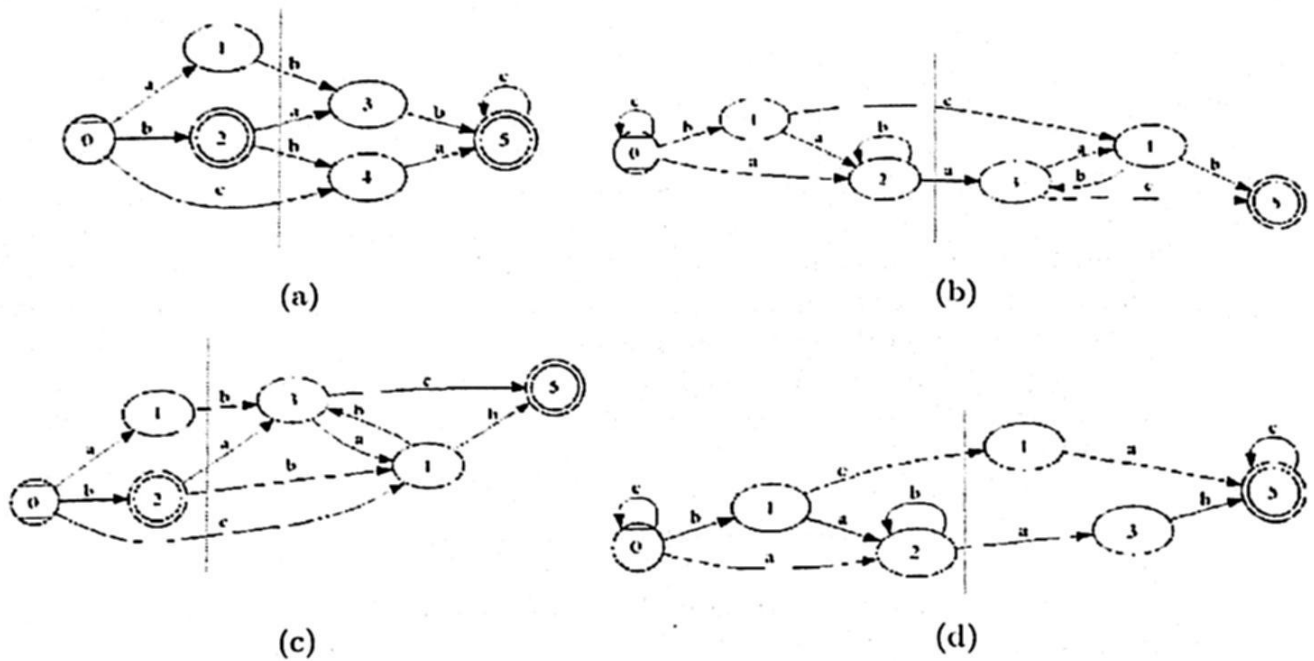


Fig. 2. Crossover between two parents: p_1 (a) and p_2 (b); the offsprings o_1 and o_2 are shown in (c) and (d), respectively

3.6 Other modules

In order to track the partial performance of the experiments, we implemented a set of additional modules which carry out the following tasks:

- **Save the main automaton:** Save the structure of the main automaton in a file for future use.
- **Get a main automaton:** Load the structure of a main automaton from a file.
- **Generate only negative samples:** This procedure overwrites the current negative samples with new ones.
- **Save positive and negative samples:** All string samples, currently in memory, are saved into a file.
- **Load samples:** It loads the string samples from a file.
- **Generate prefix tree automaton:** This module allows to obtain the prefix tree automaton from a set of positive samples.
- **Generate a *dot* file:** This module is used to generate a graphical view of one automaton (*gif* file format).
- **Generate a *SF* file:** This module is used to generate a writeout of one automaton in Standard Format (SF). The format used is the one defined for automata manipulation files in [5].
- **Replace main automaton:** Sometimes an automaton from the additional automata is required to replace the main automaton. This module offers this possibility.
- **Modify parameters:** This module allows to view and/or modify the current parameters of the system.

4 Experimental results

We have carried out different experiments in order to evaluate the performance of the proposed approach. It is obvious, that results strongly rely on the characteristics of the main automaton: the more complicated the structure is, the more difficult of inferring its grammar it will be. In our experiments we have set main automaton generation parameters as follows: the number of states was equal to 20, $ARCS_THRESHOLD=0.9$, $END_THRESHOLD=0.9$ and, $NUM_SYMBOLS=2$.

We have analysed the influence of different parameters on the system performance: for instance, the number of training samples, the size of an automata set, and the number of states of the training set. In Table 1, the obtained accuracy is shown as the number of samples grows up from 100 to 500. The size of the automata set is equal to 100 and, each automaton has 7 states. Since the results depend on the automata population, we have carried out 10 runs to exclude influence of different automata sets and then, an arithmetic mean has been calculated. In order to analyse whether or not an automata subset approximate better than the best automaton of the population (i.e., the automaton which has the highest score), we have compared accuracy obtained in both cases.

attempt	number of samples									
	100		200		300		400		500	
	setofautom.	bestautom.	setofautom.	bestautom.	setofautom.	bestautom.	setofautom.	bestautom.	setofautom.	bestautom.
1	0,950	0,855	0,955	0,950	0,872	0,842	0,955	0,974	0,792	0,786
2	0,540	0,845	0,963	0,963	0,698	0,735	0,900	0,875	0,959	0,959
3	0,885	0,955	0,685	0,885	0,955	0,952	0,864	0,799	0,932	0,886
4	0,795	0,925	0,895	0,953	0,965	0,960	0,738	0,896	0,816	0,955
5	0,745	0,955	0,950	0,950	0,955	0,947	0,931	0,940	0,847	0,870
6	0,940	0,945	0,890	0,930	0,722	0,748	0,986	0,974	0,931	0,987
7	0,880	0,890	0,808	0,780	0,692	0,837	0,918	0,974	0,974	0,959
8	0,965	0,945	0,638	0,718	0,845	0,955	0,914	0,855	0,929	0,981
9	0,515	0,735	0,950	0,950	0,952	0,955	0,726	0,868	0,961	0,970
10	0,855	0,860	0,858	0,868	0,793	0,955	0,925	0,930	0,869	0,830
Arithmetic mean	0,807	0,891	0,859	0,895	0,845	0,889	0,886	0,908	0,901	0,918

Table 1. Performance analysis of the system over different number of samples (the number of states of the automata is equal to 7)

It may be observed a slight increasing of the accuracy as the number of samples grows up, though a low accuracy may appear for any number of samples. It seems that when we generate a good set of automata (which is able to infer the grammar), a better number of samples will help to improve the results. Obviously, if the major automata have low scores (close to zero), then the increasing of the number of samples will not affect the accuracy value. Moreover, it may be seen that the best automaton of a population does not always reaches the higher accuracy with respect to the complete set. Even if a population does not have automata with high scores, a set of automata may achieve good results.

The aim of the next experiment was to investigate the influence of the number of states of the automata set with respect to its capability of obtaining the correct grammar. Thus, we have increased the number of states from 7 to 12. The results of this experiment are shown in Table 2. As expected, by increasing the number

of automata states we highly improved the accuracy and, therefore, there is no occurrence of very low accuracy, as we have seen in the previous experiment. Besides, a probability of obtaining an automaton similar to the main one also increases because the accuracy of the best automaton is generally higher than that of the complete set. Moreover, by using 16 instead of 12 states for the automata population, we may observe a similar behaviour (see Table 3).

attempt	number of samples									
	100		200		300		400		500	
	set of autom.	best autom.	set of autom.	best autom.	set of autom.	best autom.	set of autom.	best autom.	set of autom.	best autom.
1	0,760	0,935	0,940	0,943	0,945	0,930	0,884	0,898	0,951	0,952
2	0,920	0,920	0,910	0,945	0,978	0,990	0,858	0,894	0,944	0,947
3	0,785	0,900	0,943	0,943	0,933	0,960	0,959	0,978	0,910	0,948
4	0,880	0,925	0,925	0,943	0,928	0,960	0,955	0,963	0,967	0,971
5	0,930	0,920	0,765	0,948	0,963	0,947	0,936	0,950	0,949	0,920
6	0,945	0,915	0,938	0,965	0,953	0,958	0,721	0,955	0,949	0,964
7	0,850	0,920	0,943	0,943	0,930	0,958	0,934	0,955	0,948	0,947
8	0,920	0,915	0,960	0,950	0,958	0,957	0,808	0,954	0,936	0,948
9	0,935	0,935	0,915	0,975	0,965	0,975	0,873	0,946	0,946	0,948
10	0,880	0,925	0,950	0,970	0,958	0,965	0,849	0,873	0,955	0,953
Arithmetic mean	0,881	0,921	0,919	0,952	0,951	0,960	0,878	0,936	0,946	0,950

Table 2. Performance analysis varying the number of samples (the number of states of each automaton is equal to 12)

attempt	number of samples									
	100		200		300		400		500	
	set of autom.	best autom.	set of autom.	best autom.	set of autom.	best autom.	set of autom.	best autom.	set of autom.	best autom.
1	0,795	0,985	0,963	0,960	0,973	0,985	0,956	0,973	0,928	0,986
2	0,780	0,980	0,963	0,958	0,952	0,977	0,964	0,981	0,847	0,935
3	0,960	0,985	0,828	0,968	0,962	0,953	0,948	0,956	0,942	0,967
4	0,910	0,985	0,963	0,968	0,978	0,967	0,974	0,976	0,961	0,952
5	0,780	0,940	0,955	0,960	0,962	0,962	0,961	0,965	0,951	0,955
6	0,770	0,980	0,928	0,955	0,988	0,982	0,953	0,956	0,920	0,973
7	0,730	0,985	0,865	0,955	0,967	0,967	0,953	0,973	0,942	0,950
8	0,825	0,990	0,920	0,955	0,970	0,965	0,784	0,958	0,951	0,952
9	0,775	0,985	0,943	0,960	0,712	0,972	0,955	0,958	0,950	0,976
10	0,985	0,985	0,948	0,973	0,922	0,975	0,970	0,993	0,910	0,951
Arithmetic mean	0,831	0,980	0,927	0,961	0,939	0,970	0,942	0,969	0,930	0,960

Table 3. Performance analysis of the system with different number of samples (the number of states of each automaton is equal to 16)

In order to increase the probability of obtaining the best set of automata, the size of the population was also increased (see Table 4). Normally, in the grammatical inference task we lack of information about the number of states of the automaton we would infer, therefore, it is desirable to approximate it with automata containing as low number of states as possible. That is the reason because we have set the states number equal to 7 in this experiment, as well as we did in the first experiment. The size of the sample set has been chosen equal to 500 in order to evaluate better the scores of an automata population.

In Table 4, we may observe the high improvement of the accuracy and, therefore, the improvement of the automata set properties with the increase of its size. Moreover, we may see that the best automaton of the population does not usually

attempt	number of samples									
	100		200		300		400		500	
1	setofautom.	bestautom.	setofautom.	bestautom.	setofautom.	bestautom.	setofautom.	bestautom.	setofautom.	bestautom.
2	0,889	0,832	0,956	0,912	0,900	0,890	0,948	0,957	0,965	0,979
3	0,973	0,959	0,906	0,871	0,957	0,954	0,959	0,957	0,908	0,964
4	0,958	0,958	0,968	0,965	0,957	0,957	0,948	0,872	0,957	0,957
5	0,736	0,729	0,966	0,957	0,965	0,980	0,955	0,960	0,957	0,967
6	0,759	0,723	0,882	0,983	0,916	0,931	0,957	0,957	0,970	0,957
7	0,792	0,768	0,960	0,957	0,937	0,961	0,963	0,957	0,959	0,984
8	0,908	0,926	0,930	0,971	0,957	0,960	0,978	0,985	0,950	0,982
9	0,941	0,946	0,907	0,955	0,957	0,957	0,956	0,957	0,956	0,957
10	0,911	0,941	0,967	0,969	0,929	0,957	0,978	0,957	0,967	0,958
Arithmetic mean	0,777	0,790	0,902	0,957	0,917	0,957	0,947	0,964	0,905	0,970
	0,864	0,857	0,934	0,950	0,939	0,950	0,959	0,952	0,949	0,968

Table 4. Performance analysis of the system with different sizes of the automata set (number of states of the automata is equal to 7, size of the sample set is equal to 500)

achieves the highest accuracy. Thus, there still exist the problem of instability of the obtained results.

We would propose two different manners of dealing with this particular problem. On the one hand, the most optimistic one is to improve properties of the best population automaton. This process may be done by applying the genetic algorithm to an automata population. The results are very promising, even though the initial automata set generated by the proposed system is not appropriate to approximate the main automaton. After the application of the genetic algorithm we highly improved the properties of the automata and, therefore, it is obtained a high accuracy. In Table 5 the accuracy before and after the use of the genetic algorithm is presented.

	automata set	best automaton
before	0,785	0,792
after	0,970	0,970

Table 5. Comparison of results before and after the application of the genetic algorithm

On the other hand, we may also improve the system performance by choosing the best subset of the population automata; although in this case the automata characteristics remain the same and again we have a high dependence with the automata set generated. We may split the training sample set by two and, thereafter, to use the first in order to obtain learning automata scores, whereas the second part may be used in order to find the best automata set. Unfortunately, we have not performed these experiments yet, because it may be deduced that they will not improve the results of the genetic algorithm. In fact, with $NUM_AUTOMATA=500$, $NUM_POS_SAMPLES=NUM_NEG_SAMPLES=500$ and $NUM_STATES=7$, the proposed system normally achieves an accuracy close to 95%, without the application of any other additional method (see Table 4).

5 Conclusions

We have built a grammatical inference system based on the use of genetic algorithms that has shown good results. The aim of the proposed system was to infer the grammar given a finite and non-deterministic automaton, by using a set of additional automata. This automata set is automatically and randomly generated by just establishing a set of parameters in the evaluation procedure. The possibility of choosing either the best automaton or any subset of automata whose scores are higher than a given threshold for the evaluation of the test set is provided.

We have carried out different experiments in order to analyse the influence of the system parameters on its performance. There has been shown that the results strongly depend on the initialization of the automata set. By increasing the number of states we observed a raising of the automata score and, therefore, the improvement of the results. Also the increase of the size of an automata set raise the probability of generating automata similar to the main one. All these conclusions have been confirmed by the experiments carried out.

Moreover, we have seen that the application of genetic algorithms may highly improve the obtained results, even if the automata set has initial low scores. We consider that the use of genetic algorithms is a good manner of dealing with this particular problem, because the intrinsic properties of algorithms of this kind conducts a wide search over very different solutions.

The use of deterministic instead of non-deterministic automata for modeling the grammar seems to be not important, but we would like to investigate this issue in the future. Finally, the application of this system to real problems will be further analysed.

References

1. M. L. Forcada: *Neural Networks: Automata and Formal Models of Computation*, Universitat d'Alacant, Dept. Llenguatges i Sistemes Informàtics, E-03071 Alacant (Spain).
2. Y. Sakakibara: *Recent advances of grammatical inference*, Theoretical Computer Science, 185:15–45, 1997.
3. L Pitt: *Inductive Inference, DFAs and Computational Complexity*, In J Siekmann (editor), *Proceedings of the International Workshop AII 89*, Lecture Notes in Artificial Intelligence 397, pages 18–44, Springer-Verlag, 1989.
4. E.M. Gold: *Language identification in the limit*, Information and Control, 10:447–474, 1967.
5. E. Vidal: *Definition of the Standard Format (v2.0)*, Grammatical inference course, <http://web.iti.upv.es/~evidal/students/doct/ig/softw/FSformatV2.0.pdf>, 2006 (in spanish).
6. J. H. Holland: *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, Ann Arbor, MI: University of Michigan Press, 1975.

Author Index

Índice de autores

Aguilar, Raúl	157	Koh-Dzul, Roberto	337
Anzures-García, Mario	279	Kolesnikova, Olga	221
Arias-Estrada, Miguel	145	Ledeneva, Yulia	233
Arroyo F., Gustavo	209	Mata, F.	85
Barron, R.	41, 53, 317	Mejía-Lavalle, Manuel	209
Basto-Díaz, Luis	337	Molinar-Solis, Jesús E.	109
Bautista-Thompson, Ernesto	109	Moo-Mena, Francisco	337
Bello, Pedro	15	Morales, Eduardo F.	209
Brito-Guevara, Roberto	109	Moreno, Marco	191
Camas, Jorge	179	Mota, Rafael	179
Capel Tuñón, Manuel I.	291	Nuño-Maganda, Marco Aurelio	145
Carbajal Hernández, José Juan	169	Omar Ariosto Niño Prieto	3
Colmenares Guillén, Luis Enrique	3	Ordaz Gutiérrez, Susana	59
Contreras, Meliza	15	Ortega-Arjona, Jorge Luis	73
Cruz, D.	261	Osorio de Jesús, Nayely	233
Cuevas, F.J.	317, 327	Pacheco, C.	245
Curi-Quintal, Fernando	337	Paderewski-Rodríguez, Patricia	279
De Antonio, Angélica	157	Padilla, Alejandro	27
De Ita, Guillermo	15	Peña, Adriana	157
Díaz, Elva	27	Pérez, Madain	179
Enriquez, Sergio	27	Pogrebnyak, O.	317
Felipe-Riverón, Edgardo M.	303	Ponce de León, Eunice	27
Gallegos Fuentes, Francisco Javier	59	Robles, Fabián	59
García, I.	127, 245, 261	Rodríguez, Guillermo	209
García-Hernández, René Arnulfo	233	Rosales Silva, Alberto Jorge	59
Garcilazo-Ortiz, Juan	337	Rossainz López, Mario	291
Gelbukh, Alexander	221	Sánchez Fernández, Luis Pastor	97
Gomez, L.E.	41, 317, 328	Sánchez-Gálvez, Luz A.	279
Gutierrez Aldana, Alfonso	97	Sossa, J.H.	41, 317, 328
Guzmán, Giovanni	53	Suárez López, Mauricio	169
Hernández, Héctor	179	Tapia Moreno, Francisco Javier	119
Hernández-Quiroz, Francisco	73	Torres, Miguel	191
Herrera A.	127	Torres-Huitzil, Cesar	145
Horna, Luis	53	Valero Cruz, Raúl A.	169
Hornos, Miguel J.	279	Vazquez-Ferreya, Anabel	233
Imbert, Ricardo	157	Villa Martínez, Héctor Antonio	119
Isaza, Claudia	179	Villafuerte Ramírez, Miguel Santiago	97
Jimenez, J.F.	41, 317, 328	Villalobos-Castaldi, Fabiola M.	303
Juárez, Nicolás	179	Zepeda, S.	85

Editorial Board of the Volume

Comité editorial de volumen

Ajith Abraham

Juan-Manuel Ahuactzin

Arnulfo Ambler

Fernando Arango

Bedrich Benes

Miguel Martínez Rosales

Nieves BrisboaKlaus Brnnstein

Antonella Carbonaro

Andre Carvalho

Gabriel Ciobanu

Antonio Alarcón

Marco A. Moreno Armendáriz

Juan J. Flores

Ren Fuji

Frederico Fonseca

Alexander Gelbukh

Enrique Herrera Viedma

Grigori Sidorov

Mario Aldape

Abel Gómez

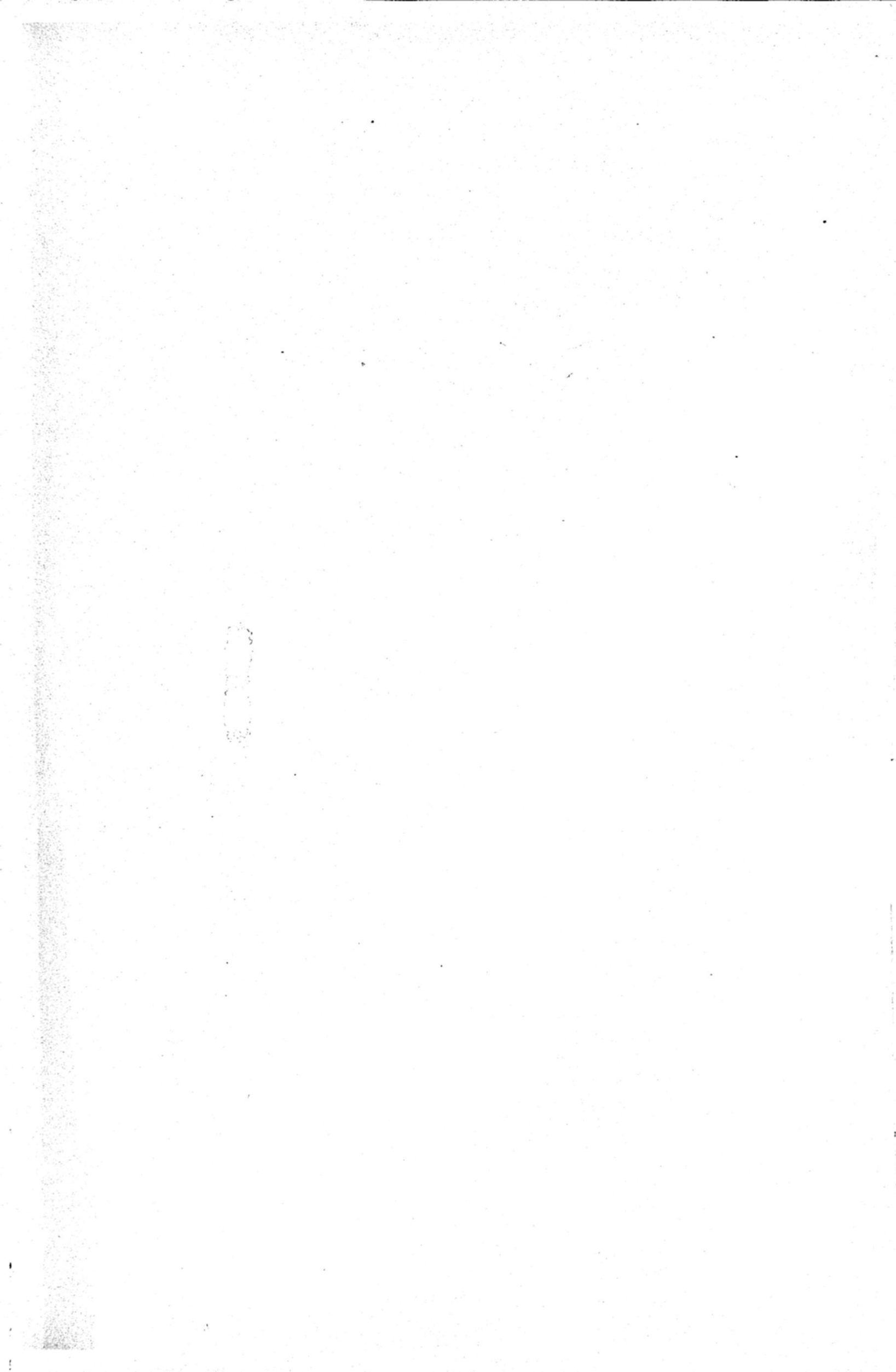
Israel Román

Yulia Ledeneva

Olga Kolesnikova

Itzamá López

**Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
Mayo de 2010
Printing 500/ Edición 500 ejemplares**



This volume contains 21 carefully selected papers by 61 authors from 10 countries: Cameroon, Canada, France, Iceland, India, South Korea, Mexico, Portugal, UK, and USA. These papers present the most recent developments in a range of areas related to computer science and engineering. The papers are arranged into 10 thematic fields:

- Logic Programming
- Optimization and Classification
- Neural Networks and Evolutionary Algorithms
- Bioinformatics and Medical Applications
- Software Engineering
- Cryptography and Security
- Computer Networks
- Educational Software
- Control
- Computer Architecture

The volume will be useful for researchers, students, and general public interested in the corresponding areas of computer science and engineering.

ISSN: 1870-4069

www.ipn.mx

www.cic.ipn.mx



INSTITUTO POLITÉCNICO NACIONAL
"La Técnica al Servicio de la Patria"

